

Climate debate must not overheat

Charges by parts of the US energy industry that a recent report on global climate change has been 'scientifically cleansed' should not be allowed to undermine efforts to win political support for abatement strategies.

ONE of the most significant implications of the forthcoming presidential and congressional elections in the United States will be their impact on international negotiations about the dangers of man-made global warming. Despite continued dissent from a dwindling band of sceptics, there is growing support within the scientific community for the view, expressed in the latest report from the International Panel on Climate Change (IPCC) (see page 546), that "the balance of evidence suggests a discernible human influence on global climate". The main political task now facing the international community is to turn this conclusion into a plan of action. The extent to which the United States is able to commit itself to this task will be crucial to its success.

There is, therefore, good reason to be concerned about the danger that US politicians may be deflected from this goal by spurious attempts to undermine the validity of the IPCC's conclusions. In the vanguard of such efforts has for several years been the Global Climate Coalition, a group whose description as "an organization of business trade associations and private companies" disguises the extent to which it acts, in particular, as the voice of fossil-fuel producers, primarily in the oil, coal and petroleum industry. Many such bodies are justifiably concerned that attempts to limit the emission of greenhouse gases will impinge on their future activities.

Last week, the GCC published a strongly worded criticism of the IPCC's latest report which, it claimed, had been subject to an institutionalized "scientific cleansing". The pressure group's main target is a chapter in the report of IPCC's Working Group 1 — the panel responsible for assessing the science of climate change — on the detection of such change and the attribution of causes. The GCC's main complaint is that editorial changes were made between the formal approval of this chapter by the full working group at a meeting in Madrid last November and its subsequent appearance in print.

The complaints are not entirely groundless. IPCC officials claim that the sole reason for the revisions was to tidy up the text, and in particular to ensure that it conformed to a 'policymakers' summary' of the full report that was tortuously agreed by government delegates at the Madrid meeting. But there is some evidence that the revision process did result in a subtle shift in the relative weight given to different types of arguments, and that — not surprisingly — this shift tended to favour arguments that aligned with the report's broad conclusions. Conversely, some phrases that might have been (mis)interpreted as undermining these conclusions, particularly if, as IPCC officials feared, they were taken out of context, have disappeared.

The GCC has chosen to express its concern on the question of whether such changes are within the rules of the IPCC. It is certainly an issue that deserves attention. IPCC officials insist that the organization's rules allow changes to be made after draft documents have been accepted in principle by the full working groups. But on a topic where political sensitivities run high, the integrity of the reviewing and approval process is, as in any scientific publishing endeavour, an essential element in assuring the credibility of the resulting conclusions. Public clarification on this point would be useful.

Having said that, however, it is also essential that procedural criticism is not allowed to throw out the baby with the bathwater.

The IPCC's Working Group 1 has taken pains, as its full report indicates, to meet scientific criticisms of its conclusions head on. Inevitably there are those who will continue to dispute the details of its arguments, and others who will be reluctant to sign up to the consensus for reasons of personal caution. Yet the GCC's attack on the IPCC allows direct consideration of the science involved to be overshadowed by semantic arguments that risk losing the wood in the trees. Climate change as a political issue deserves the increasing attention it is likely to obtain as governments in Washington and elsewhere digest the IPCC's conclusions. But it also deserves a more thoughtful and constructive assessment than one that merely seeks to impugn the motives of scientific participants by alleging that they are responding to political pressures. □

Replacing Cluster

European space scientists learnt a hard lesson last week: there is no such thing as a free launch.

TEARS flowed last week along with the burning debris of Cluster's four satellites and Ariane-5's first launcher, which aborted its mission soon after lift-off (see page 541). The first big failure of the European Space Agency (ESA)'s science programme raised an immediate question: should the space agency have risked the price of the whole mission by trusting such a hugely expensive and important project to an untested launcher, just because it was offered a free ride? The answer is yes. ESA officials maintain that the risk of an Ariane-5 launch failure was no greater than that of any other launcher capable of handling the 5-tonne load. They would make the same decision again, they say. And on top of this, the hundreds of millions of ECUs in savings, at a time of general economic strain, cannot be sneezed at.

The real question now is how the scientific community should be compensated for the loss. It is out of the question that the ECU500-million mission should simply be repeated, even though a second attempt would be much cheaper. But it is equally out of the question that Cluster should be written off entirely. The ESA science programme is designed to cover the needs of all space scientists in Europe, and the solar-terrestrial research community cannot be left empty-handed. ESA also has a wider responsibility, as Cluster was a significant element in the coordinated international effort called the Interagency Solar Terrestrial Programme, along with missions from the space agencies of the United States, Japan and Russia.

The question is: what sort of compensatory mission should now be offered, and who should pay for it? Although ESA's space science directorate must be congratulated for tackling this question speedily, money from its own overstretched budget can be given only at some cost to its other planned space science missions. But there is another option. The Cluster/Ariane-5 disaster is a setback for the whole agency, not just its science directorate. The director general of ESA, Jean-Marie Luton, should recognize this, and consider asking the more financially well-endowed launcher directorate to help pay for a new, smaller mission. □



ESA seeks to pick up the pieces from Cluster mission's fiery fate

Munich. The European Space Agency (ESA) is to decide within a few weeks whether partially to replace its space science mission Cluster, a group of four satellites planned to explore the 'solar wind' which was lost when the launcher, Ariane-5, exploded during its maiden flight last week.

But even if a scaled-down repeat mission is approved, there is no guarantee that Cluster's unique three-dimensional imaging capability will survive. Nor is there any guarantee that national space agencies will finance new payloads.

Cluster was one half of ESA's ECU-850-million (US\$1.05-billion) solar-terrestrial physics programme, the first of ESA's four 'cornerstone' missions of its Horizon 2000 space science programme. The other half of the mission is SOHO, which was successfully launched a few months ago, and will image the Sun's disk, corona and wind at different wavelengths.

Cluster was planned to study the charged particle, electrical and magnetic field environment of the Earth in its responses to solar activity, as monitored by SOHO. Its four identical satellites would have provided the first high-resolution, three-dimensional

analysis of this environment. SOHO-Cluster is also part of a wider global programme of solar-terrestrial physics, the Interagency Solar Terrestrial Programme (ISTP). This is a series of coordinated missions that include NASA's Wind, Japan's Geotail and Russia's soon-to-be-launched Interbol.

Roger Bonnet, head of ESA's space

reliable option at the time, with the technical capability of lifting the five-tonne weight of the four-satellite mission. "I would take the same decision again," he says.

Scientists are aware of the relatively high failure rate of all launchers; that of Ariane-4, for example, is one in twelve. But this did not help ease the pain felt by Cluster investigators last week, who watched years of their working lives disintegrate 40 seconds after launch. "I realized it could happen," says André Balogh, of Imperial College, London. "But I had given no structured thought to what would happen in the eventuality."

In fact, ESA has been quick to move on behalf of the bereft scientists, and a plan of action is likely to be agreed in the coming weeks. ESA feels a great responsibility to the scientists who lost so much, says Bonnet, pointing to their "total despair, depression and tears" immediately after the explosion.

Bonnet has already commissioned from Cluster scientists an inventory of their remaining hardware, software, manpower and facilities, in order to assess how far the original scientific objectives of Cluster might be salvaged. ESA's science advisory group and its science programme committee (SPC), which is made up of representatives of its member states, will meet next week to explore how to proceed.

There is general agreement in the space science community that the Cluster mission cannot be fully replaced because of a shortage of money in the space science budget. But there are also very strong feelings that something of the mission should be saved. "Given the scale of the loss, ESA should feel obliged to compensate [the scientific community] in some way," says Berend Wilken, an investigator from Germany's Max Planck Institute for Aeronomics, reflecting the general mood of the 900 or so scientists involved in Cluster.

Sympathy abounds both within and around ESA. "It is hard to imagine how we could get the whole mission back," says David Southwood, chairman of the SPC. "But we will trawl for clever ideas of what we can do with what is available." Bonnet points out that the ESA space science programme "is intended to serve all space science communities and solar terrestrial scientists must be left with something".

But he also emphasized that no other mission within the space science programme will be sacrificed to make room for a Cluster replacement. The only financial ►

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Exit Ariane-5: explosion devastated Cluster scientists.

science directorate, defends the decision made by ESA in 1986 to accept an offer of a free flight on Ariane-5's first qualification launch, not only because this saved hundreds of millions of ECUs, but also because the launcher was thought to be the most

NASA tightens criteria for new launchers

Washington. The US National Aeronautics and Space Administration (NASA), which has a backlog of scientific satellites awaiting launch because of a string of recent rocket failures (see *Nature* 376, 717; 1995), is to increase its supervision of untested launch vehicles, and will require them to institute new quality control measures. The requirements will be included in a request for proposals for small launch vehicles scheduled to go out to industry this summer.

Problems with the new Pegasus XL launch vehicle built by Orbital Sciences Corporation of Fairfax, Virginia, have been particularly vexing to NASA. Four of the agency's small science satellites have been on hold because of the rocket's failure in its first two outings. Now that the Pegasus is back in service (with a success in March), all four are scheduled for launch this year. Another small launcher built by Lockheed Martin Corporation, which also failed on its first try, is scheduled to carry NASA's Lunar Prospector in 1997.

Frustrated by the delays, Daniel Goldin, the administrator of NASA, announced last year that no NASA spacecraft would ride on a first-time launcher in the future. That

policy will still hold for unique or expensive missions where the risk of failure is high. But some lower-risk payloads may still end up on maiden flights, says Karen Poniatowski of the agency's launch vehicles office.

The new NASA policy divides launchers into three categories: those with no proven flights, those with some experience (1 to 14 flights), and mature launch systems. Launchers in all three categories would have to submit to an ISO 9000 review — an international quality review standard — before a launch contract could be awarded.

Launchers in the first two categories would be subject to several additional reporting requirements, including a NASA audit with "relatively small on-site requirements", says Poniatowski.

So far, says Poniatowski, the launch industry seems willing to accept the new NASA policy. Although there may be some increase in costs associated with the additional requirements, the improved quality control should result in savings, she says. The intention of the new policy is to let the launch industry continue to audit itself, while boosting NASA's confidence in a successful launch.

Tony Reichhardt

► flexibility that ESA has to achieve this would be to delay future missions, while trying to avoid a detrimental knock-on effect.

What are the realistic options for an affordable, scaled-down version of Cluster? One possibility would be to fly a single satellite with simpler scientific instruments. One back-up satellite already exists and could be prepared quickly and reasonably cheaply by the German contractor, Dornier.

A second possibility would be to launch the back-up satellite along with one or more smaller, simpler and cheaper satellites. But this would require much more time, as the design of any new satellite would need to be extensively tested.

In either case, ESA will be looking at the possibility of putting any replacement satellite or satellites into orbit with other ISTP spacecraft, in an attempt to reproduce the three-dimensional resolving power characteristic of Cluster's four-satellite set-up, which scientists are keen to protect.

In both cases, the reduced weight of the mission would allow a cheaper launcher — such as Ariane-4 — to be used. In both cases, however, even if ESA was able to find sufficient money to provide the replacement satellites, national space agencies would have to agree to pay once again for the instruments, despite extreme tightness of their own funds.

Alison Abbott

NIH resists bill to promote research into Parkinson's

Washington. Support is growing in the US Congress for a bill that would dramatically increase spending on research into Parkinson's disease by the National Institutes of Health (NIH) — despite the opposition of leading NIH officials who dislike congressional mandates.

The bill, named after Morris K. Udall, a popular former congressman, has 42 Senate co-sponsors and 176 supporters in the House of Representatives, ranging from conservative Republicans to liberal Democrats — a “dream” list, according to one Senate aide. Between 1 and 1.5 million Americans are believed to suffer from Parkinson's disease. One estimate puts the cost of their care at \$25 billion a year. The bill authorizes more than \$100 million a year for Parkinson's research over five years, compared to \$28 million at present.

Advocates describe the present sum, much of which is channelled through the National Institute of Neurological Disorders and Stroke (NINDS), as paltry compared to that allocated to diseases such as Alzhei-

mer's and AIDS. They have now convinced many politicians of their case. Even conservative opponents of fetal tissue transplants — which NIH Parkinson's money now supports — have signed on as co-sponsors, even though the bill does not contain a ban on NIH funding of such research.

One such senator, Rick Santorum (Republican, Pennsylvania), has “some reservations” about the bill because of this, says his spokesman, Tony Fratto. “But we felt that [Parkinson's disease] is so important that we need to give the bill as much support as possible.” Other senators, such as Dan Coats (Republican, Indiana), who vocally oppose congressional “earmarking” of NIH funds, are making an exception and supporting the Parkinson's bill.

Zach Hall, the director of NINDS, declines to comment on the bill, but says his institute is “very actively concerned” with Parkinson's.

Harold Varmus, the director of NIH, questioned in March by the Senate Committee on Labor and Human Resources, said he opposed the bill. He argued that the longed-for advances in Parkinson's research would come about not by quadrupling funds directed at Parkinson's but from “work that is generic to all nerve cells”.

But Paul Wellstone (Democrat, Minnesota), both of whose parents had the disease, challenged Varmus, saying that many people struggling with the illness “are not at all persuaded” that generic research is enough. Citing “no increase at all” in funding for Parkinson's by NIH since 1989, Wellstone said: “I don't see that much of a commitment.”

Varmus replied that he feared the NIH could not “responsibly fund” ten Parkinson's research centres, as required by the bill. “We don't believe that science would be furthered by such a large expenditure of money for so many centers,” he said. He also called on the committee — which is responsible for drafting a separate bill that authorizes the NIH — to preserve “an atmosphere as free as possible from restrictions” requiring the NIH to spend money on certain diseases. The Udall bill was introduced in 1995. Its House author is Henry Waxman (Democrat, California), and its Senate author is Mark Hatfield (Republican, Oregon).

The bill's advocates hope that the Senate Committee on Labor and Human Resources will soon amend a bill reauthorizing the NIH to include the Udall bill. But chairwoman Nancy Kassebaum (Republican, Kansas), opposes additions to the NIH bill, and its House counterpart is unlikely to see legislative action this year, according to aides.

Meredith Wadman

Chinese rocket site ‘blind to safety’

Hong Kong. An internal memorandum from the international satellite company Intelsat, criticizing a “blindness towards safety” at China's rocket launch site and describing it as “pathetically short of world standards”, has added to the aerospace industry's growing concerns about the viability of China's commercial satellite launch capabilities.

The memorandum, which was obtained by the Washington-based publication *Science and Government Report*, reports what it claims to have been “callous disregard for human life” and states that “under no circumstances” can Intelsat use the site again. It is written by a manager who witnessed the explosion on 14 February during attempted lift-off on the first flight of a Long March 3B rocket of Intelsat's 708 satellite from the space site at Xichang in Sichuan province.

Although China's Great Wall Industry Corporation, which runs the launch facility, reported four deaths, Israeli film footage smuggled out afterwards showed soldiers clearing bodies and wreckage over a large area. The film-maker estimated that a hundred people had been killed. In another disaster in January 1995, six people were officially reported killed at the inland military site, which is in a mountainous region popu-

lated by farmers.

The Intelsat memo says observers were prevented from leaving the viewing area for nine hours after the explosion. “In retrospect, this gave the Chinese enough time to clear out any dead people from the gate and village areas, which were not evacuated to our knowledge,” the memo says.

Great Wall declines to comment on either its launch schedule or the site's safety. Since the disaster, however, two planned Intelsat launches, one by the US-based company Echostar and a fourth by Asia Satellite Communications (AsiaSat), based in Hong Kong, have been moved to Russia's Proton and France's Ariane rockets.

February's failure contrasts sharply with that of Ariane (see page 241), in which the rocket was blown up from the ground within a few seconds of launch. The Long March 3B was allowed to fly for 20 seconds and to turn until it was heading back to Earth, even though television footage clearly showed it turning off-course before it had even cleared the launch tower. Some observers speculate that either the Chinese held off to avoid damaging the launch tower or that their rocket was not carrying explosives capable of destroying the vehicle in case of difficulties.

Elisabeth Tacey

Labs collide over rival tritium schemes

Pleasanton, California. A \$2-billion prototype fusion machine has been suggested by the former director of the Lawrence Livermore National Laboratory (LLNL) in California as an alternative means of producing tritium for US nuclear weapons to the planned particle accelerator.

John Nuckolls, who pioneered inertial confinement fusion research with Edward Teller, and who was director of LLNL until 1994, says that such a machine could help to "make the transition" between the planned National Ignition Facility (NIF) — which is expected to demonstrate the feasibility of inertial confinement fusion — and a commercially viable system.

The United States says it will need a supply of tritium to replenish its nuclear weapons in about ten years, and the Los Alamos National Laboratory in New Mexico is leading a \$280-million research programme to design a particle accelerator to produce it (see *Nature* 377, 567; 1995). All US nuclear weapons contain tritium, which decays with a half-life of eleven years.

The accelerator being designed at Los Alamos would produce 3 kg of tritium a year by firing a proton beam into a tungsten target to generate neutrons, which are then absorbed in helium-3 to make the tritium. It would cost around \$2 billion to build, but would consume 400 MW of electricity — enough for a city of half a million people — and could therefore cost as much as \$200 million a year to operate.

The Department of Energy has already said that it will decide in 1998 whether to build the accelerator; if it goes ahead, it would be built at Savannah River, South Carolina (see *Nature* 376, 201; 1995).

But Nuckolls told a meeting of Fusion Power Associates, a pro-fusion lobby group, at Pleasanton, California, two weeks ago that if a START 3 treaty is agreed with Russia and the tritium requirement is delayed, an inertial confinement fusion machine could be built instead of the accelerator.

"I know it is going to be awkward explaining this to our friends at Los Alamos," Nuckolls said. "But I think we've got to look at it." Los Alamos and Livermore have been feuding about projects and prestige ever since Teller set up the California lab in 1952.

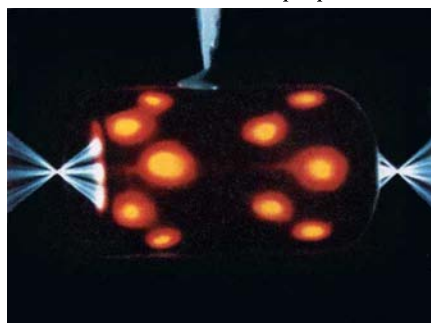
Nuckolls suggests that the prototype machine could use either a solid-state laser or a beam of heavy ions to induce fusion in tiny pellets of deuterium fuel. The pellets might be contained in a small cylinder. This converts the beams into intense X-rays, which compress the pellet to induce fusion.

He envisages a machine which would do this every two seconds, using a pulse of 5 megajoules of energy each time to induce 40 times as much energy by fusion. The laser or ion beam could be built for \$1 billion, he

argues, leaving \$100 million for a factory to produce the pellets and \$900 million for the equipment to contain the fusion blasts.

Several experts at the meeting said that most of the energy is given off as heat by the neutrons as they are absorbed, and containment is therefore not regarded as a major technical challenge.

Paul Lisowski, head of the Accelerator Production of Tritium programme at Los Alamos, said of Nuckolls' proposal: "It is



Focusing the debate: a laser target Hohlraum, which creates fusion by reflecting X-rays onto a one-millimetre fuel capsule.

possible that the [inertial confinement fusion] technology will be very productive in the future — but it is not the kind of proven technology which you could use to produce tritium for national security purposes."

The proposed 40-fold 'gain' for Nuckolls' machine would be four times as great as that expected from the \$1-billion National Ignition Facility (NIF), which is being designed at Lawrence Livermore. NIF is funded as part of the stockpile stewardship programme, which looks after nuclear weapons. But its main scientific task will be to establish that inertial confinement fusion can be induced under controlled circumstances.

Cancer genetics group drafts guidelines

Washington. Women found to have genetic mutations predisposing them to breast cancer should begin having annual mammograms between the ages of 25 and 35, according to a draft report by a US National Institutes of Health (NIH) task force due to be published soon.

This and other recommendations on the management of asymptomatic carriers of cancer-predisposing mutations result from a year-long study by a task force of the NIH Cancer Genetics Studies Consortium.

Wylie Burke, an associate professor of medicine at the University of Washington in Seattle, and chair of the task force, outlined the recommendations at a meeting last month of the Advisory Council to the National Center for Human Genome Research.

The recommendations deal with surveillance and prophylactic surgery in healthy carriers of mutations predisposing to

NIF will test out pellets which are ignited directly and indirectly, inside Hohlraums, but it will operate slowly, its optical systems taking hours to cool down after each "shot".

The Clinton administration intends to announce its final decision to build NIF at Lawrence Livermore just before November's presidential election, in which Californian votes will play a critical role.

The principles of inertial confinement fusion, including the role of the Hohlraum, were classified in the United States until 1993. Declassification has given fresh impetus to the concept, at a time when the alternative fusion technology — magnetic fusion, which involves containing a ring of hot plasma gas in a ring or tokamak — is struggling to prove its feasibility as a potential energy source.

The chief advantage of inertial confinement fusion is that, on the basis of results from classified underground tests, physicists are already confident that it will work. But it is at an early stage of development at a time when no money is available to back new energy sources. Critics also predict that an energy source based on repeated, albeit very small, thermonuclear explosions, will lack public appeal.

Nuckolls says the inertial confinement fusion community has been "unified as never before" under the stockpile stewardship programme. But he predicts that stockpile stewardship will come under funding pressure, and that defence money will be unavailable for a machine after NIF, unless it has a specific defence function such as tritium production or the destruction of plutonium.

He describes his own idea as "out-of-the-box thinking", adding that "the community will have to engage in more of that" if it is to progress after NIF.

Colin Macilwain

breast, ovarian, prostate, colon and endometrial cancers. In addition to early mammography for *BRCA1* or *BRCA2* mutation carriers, for example, the draft recommends that the former should undergo pelvic examination and transvaginal ultrasound every six to 12 months, beginning between the ages of 25 and 35, to detect ovarian cancer.

Burke says the recommendations are based on "scanty" and "lacking" information in a field greatly in need of more research. For example, there have been no studies of the effectiveness of early mammography in mutation carriers. "We are doing the best we can with the information at hand," she says. "The most important issue is to think through [current] research efforts to ensure that they produce the kind of information we were looking for and couldn't find."

Meredith Wadman

Did UK 'dump' contaminated feed after ban?

Paris. British exports of animal feedstuffs potentially contaminated with the agent that causes bovine spongiform encephalopathy (BSE) more than doubled in the years following their prohibition in ruminant feed in the United Kingdom itself, according to official government statistics.

Britain banned the inclusion of ruminant-derived protein in ruminant feed in June 1988 (although not in all meat and bone meal). In September 1990, it extended this to prohibit the inclusion of specified bovine offals (SBOs) — the most infective parts, such as brain and spinal cord — in all meat and bone meal.

But export statistics from HM Customs and Excise show that whereas UK exports of animal feeds had remained almost constant in the years leading up to the 1988 ban, they more than doubled the next year (see figure, below). Most of this increase was accounted for by exports to France.

Some of the feed exported before the SBO ban came into force would have been contaminated. Moreover, although Britain banned the export of SBOs at the same time as it introduced the SBO ban within the country, even feedstuffs exported in later years may have been contaminated, given that it is now acknowledged that the ban was not properly enforced.

Nonetheless, many veterinary experts defend the UK's decision to continue exporting feed. One member of the UK Spongiform Encephalopathy Advisory Committee, for example, argues that while the 1988 ban prohibited the inclusion of ruminant protein in rations for ruminants, it allowed the use of meat and bone meal in pigs and poultry rations. There was therefore no reason not to export it, he says.

Some experts in other countries, however, disagree with the UK's action. "They knew at that time that meat and bone meal was dangerous, yet they exported it and spread the danger of new cases of BSE arising in member states," says Udo Weimer, an official at the German agriculture ministry's animal diseases division. He adds that when

imports were banned by the member states of the European Union, the United Kingdom continued to export feed to other countries.

An official from the British Veterinary Association says that he warned the government at the time of the dangers of exporting feed, but that his warnings were ignored. "I badgered our chief veterinary officer, saying that having identified a 'poisoned food' it was immoral to export it," he claims. But, he adds: "I was firmly put in my place, and told that it was up to the importing countries to put in place all the guarantees needed."

The official also claims that a Dutch veterinary surgeon working at the port of Rotterdam told him at the time that he was "absolutely certain" that British feed was being dumped, and also speculated that some may have been rebagged and reimported into the United Kingdom.

Whether there was dumping or not, exports were eagerly snapped up because they were cheap, says Mark Savay, professor of pathology at the Centre Nationale d'Etudes Vétérinaires et Alimentaires at Maisons-Alfort near Paris, adding that the British ban had caused prices of UK meat and bone meal to plummet.

One expert agrees that feed exported before the SBO ban would have been contaminated. But he argues that the vast majority of imports must have been used for pig and poultry rations, as its toxicity would otherwise have resulted in more cases of BSE in continental Europe, whereas only around 400 cases have been identified (compared with a total of around 160,000 in the United Kingdom).

Savay argues, however, that the impact of the imports of "contaminated feed" on BSE levels in France could have been much worse if France had not traditionally used lower concentrations of meat and bone meal in cattle feed, typically around 1.5 per cent. High-protein feeds were mainly used

for pigs and poultry, he adds. (Indeed, one factor involved in the spread of BSE in the United Kingdom was that inclusion of meat and bone meal in animal feeds jumped from 1 to 12 per cent during the 1980s, a shift that resulted from a fall in the value of the pound and a corresponding increase in the cost of soya and fish meal.)

The French ministry of agriculture says that it was powerless to prevent imports of UK meat and bone meal, given that this needed to circulate freely

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Eager feeders: feed exports more than doubled in 1989.

within the then European Economic Community. But it points out that, as an intermediate measure, it prohibited imports of UK meal in August 1989 unless this was destined for animals other than ruminants. "But it would have been much better if the United Kingdom had not exported it in the first place," says a ministry official.

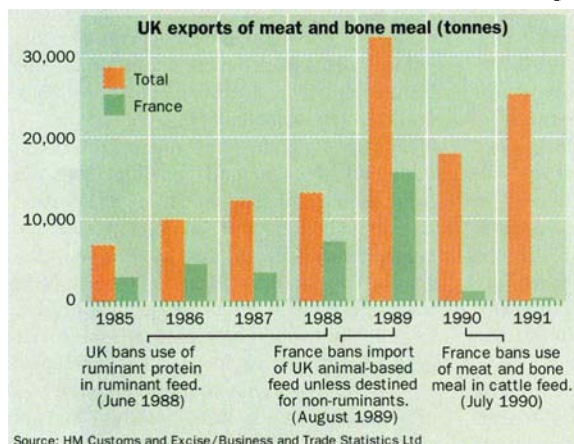
At the time, the European Commission felt it was unnecessary to apply a Community-wide ban on UK feeds, says an official from its agriculture directorate. One reason, he claims, is that the United Kingdom had given the commission assurances that it would not export feed — an assurance that was not respected, he adds — and member states had implemented their own import bans. (Germany, for example, banned imports of UK meat and bone meal in 1990.)

Another reason was that the commission felt it would be difficult to enforce a ban, he says, as feed is not labelled precisely. Even now, EU law requires feed to be labelled only as containing "products of animal origin" and not to provide further details. "Industry has never been keen on telling farmers what it puts in rations," he says, adding that there is some justification in that contents are adjusted continuously on the basis of the costs of ingredients.

Indeed, one problem in interpreting export statistics is that these group all feeds containing animal matter under the heading "flours, meals and pellets of meat of meat offal; greaves". They make no distinction, for example, between feed prepared from poultry or from cattle.

It is thus impossible to tell whether UK exports respected the 1990 UK ban on the export of feeds containing specified bovine offals. The statistics show, however, that as member states in the EU began to impose their own import bans on UK meat and bone meal around 1990, UK exports continued to grow through increased sales to countries outside the EU.

In 1991, for example, Israel imported almost 10,000 tonnes of UK feed and Thailand 6,200 tonnes, both up from zero a few years previously. Savay says that he



would like assurances that such exports did not contain specified bovine offals. "Until we know the exact nature of the exports to Thailand and other countries and what became of them, there exists the risk that they may have been given to cows and caused further contamination," he says, pointing out that many countries still lack surveillance networks for BSE.

For his part, A. Shimshony, Israel's director of Veterinary Services and Animal Health, is adamant that the imports into Israel consisted only of meat and bone meal derived from poultry. He points out that Israel banned the import of all mammalian meat and bone meal for feed use in 1990.

Moreover, imports of meal derived from poultry were permitted only from mills certified to have never produced feeds of other animal origin, he says, in order to eliminate the risk of cross-contamination. Indeed, Israel has a BSE surveillance network and has not detected any cases.

Johnston says that Israel's vigilance would be the exception rather than the rule, given that meat and bone meal is produced from cadavers, bones and other wastes from a variety of animals, and that it is consequently often difficult to distinguish what is and what is not derived from ruminants. Half a gram of infected meal is enough to kill a cow, he adds.

Indeed, Israel's action in 1990 seems almost clairvoyant. Most feed mills use the same equipment for producing all feeds. As a result, contaminated material intended for use in pig and poultry rations in the United Kingdom may have found its way into cattle feed. This is now acknowledged probably to account for most of the 27,100 cases of BSE in Britain since the introduction of the 1988 feed ban. But it was only in March this year that Britain recognized the risk and imposed a comprehensive ban on the feeding of mammalian-derived meat and bone meal to all farm animals.

While the recriminations over UK feed exports are likely to continue, the mad cow crisis has sent a shockwave through the feed industry that bodes well for the future safety of feed. The EU's council of agriculture ministers agreed in April that by the end of the year all rendering plants must convert to batch processing methods that use a temperature of 133 °C and pressure of 3 bars for at least 20 minutes. This means that "other epidemics of this nature cannot occur again", says Weimer.

Meanwhile, feedstuffs are not the only controversial exports from the United Kingdom in the BSE crisis. Many European countries are remaining remarkably discreet about the fact that they have also imported large numbers of breeding animals from BSE-infected herds in the United Kingdom, many of which could be incubating the disease. This means that although the Netherlands, for example, is listed as BSE-free, its disease status would probably be better described as unknown.

Declan Butler

Science agencies benefit as Congress eases up on cuts

Washington. The US Congress is quietly retreating from its previously stated objective of slashing the federal budget this year, leaving most science agencies in better shape than they were only a few months ago.

Bloodied by last year's prolonged budget conflict — from which it wrested billions of dollars of spending cuts but no political credit — the Republican-controlled Congress is now rushing to complete spending bills which will be more generous than expected, postponing tough action to balance the budget.

The House of Representatives, for example, is expected to complete each of the thirteen appropriations bills needed for the 1997 financial year, which begins 1 October, by early next month.

Next week, the Energy and Water appropriations subcommittee, chaired by John Myers (Republican, Indiana), will mark up a bill requiring cuts of around \$200 million. Earlier in the year, Myers was asked to make cuts of \$1 billion. His refusal to do so helped force an agreement last week between the House and the Senate which freed up an extra \$4 billion across the government.

The first bill of major importance to science funding was approved by the Veterans Affairs, Housing and Urban Development and independent agencies (VA-HUD)

subcommittee two weeks ago. The subcommittee approved a budget for NASA of \$13.6 billion, compared with \$13.8 billion requested by President Bill Clinton. The \$200 million was cut from the Mission to Planet Earth programme. Science programmes at the Environmental Protection Agency would receive \$540 million, against \$580 million requested by Clinton.

Also in the VA-HUD bill, the National Science Foundation would get \$3.25 billion, against \$3.325 billion requested by Clinton. But the subcommittee was allocated an additional \$350 million under the new House-Senate agreement, and science lobbyists are working to get some of it for the National Science Foundation.

This week, the Labor, Health and Human Services and Education subcommittee was expected to mark up a bill that will contain more money for the National Institutes of Health (NIH) than the \$12.4 billion requested by Clinton. The president's proposal would have allocated most of the NIH's 4 per cent budget increase to the construction of a new clinical centre (see *Nature* **380**, 187; 1996). But John Porter (Republican, Illinois), the chairman of the subcommittee, is expected to secure a 6.5 per cent increase in NIH research grants, as well as money to start the new centre, in his bill.

Colin Macilwain

European vote raises bioethics stakes

Munich. The European Parliament is being asked to vote later this month on a resolution opposing all research on human embryos and on people who cannot give informed consent — referred to as "legally incapacitated persons" — even though such research is currently allowed in several of the 15 member states of the European Union that the parliament represents.

Such a vote would have no legal status. But it would be intended to influence a draft bioethics convention being drawn up by the Council of Europe, a body representing 39 European countries, one of whose aims is to defend human rights. The new draft was approved last week by the council's executive, and will be voted on at its parliamentary assembly in September.

An initial draft of the convention was approved last year by this assembly by a narrow majority, on the understanding that it would undergo extensive rewording aimed at finding a consensus between countries such as Germany, which would like tough restrictions on

research on humans and embryos, and France and Britain, which want more flexibility.

The new draft tightens up the wording, making explicit, for example, that the creation of human embryos for research purposes should be prohibited. But it accepts the principle that embryo research is otherwise allowable. It also allows research in principle on legally incapacitated persons.

Responding to an initiative of the European Parliament's research committee, headed by Christof Tannert, a member of Germany's Social Democrats, the legal affairs committee of the parliament last week approved a position paper requesting a long list of amendments to the draft.

Individual members of the Council of Europe will not be required to ratify the bioethics convention, even if the new draft is approved in September. But it will provide a form of indirect political pressure on countries such as Germany if it appears to have widespread approval.

Alison Abbott

Climate report 'subject to scientific cleansing'

London. A US lobby group partly financed by oil, power and automobile companies is trying to undermine the credibility of the latest report of the UN Intergovernmental Panel on Climate Change (IPCC) by accusing it of "scientific cleansing".

The accusation has been made by the Global Climate Coalition (GCC), an umbrella group of 60 industrial concerns, in a document that is currently being widely circulated in the Congress and elsewhere in Washington DC.

Its target is the latest five-yearly report of the IPCC, agreed in principle last November and published in London last week, which concludes in particular that the balance of scientific evidence "suggests a discernible human influence on global climate".

Such a conclusion is likely to boost demands for greater curbs on the use of fossil fuels. The coalition is now trying to throw doubt on its validity on procedural grounds, in particular by criticizing the decision of one of the report's authors to re-edit a chapter before final publication, after it had been peer-reviewed and approved.

But Ben Santer, the author concerned, an atmospheric scientist at the Lawrence Livermore National Laboratory in California, says the changes were endorsed by the IPCC and were necessary "to improve the report's scientific clarity". He describes the GCC's allegations as "dangerous and absurd".

The GCC last week issued a nine-page, line-by-line analysis of the changes made to Chapter eight of the report, dealing with the

IMAGE
UNAVAILABLE
FOR COPYRIGHT
REASONS

Sign of the times? Drought in the US.

question of potential human influence on climate change. John Shlaes, executive director of the GCC, argues that, under IPCC rules, any extra editing should have been peer-reviewed.

Shlaes concludes that the new version is unbalanced, contains "substantial deletions and significant changes" to the scientific material, and over-emphasizes the role of human activities in climate change.

He has written to Bert Bolin, emeritus professor of meteorology at the University of Stockholm and chairman of the IPCC, asking him to justify the decision to authorize additional editing. Copies of the letter have been sent to senior policy-makers in both houses of Congress.

The GCC analysis claims, for example, that the concluding summary has been deleted from the final report, and that a paragraph towards the end of the chapter plays down the uncertainty of establishing and forecasting human-induced climate change.

Pat Carter/Liaison

"These revisions raise very serious questions about whether the IPCC has compromised, or even lost, its scientific credibility," says Shlaes. "The changes are not just about grammar and punctuation. They go deeper and we want to know why they were made."

But Santer defends the changes, which, he says, "can all be scientifically justified" and which were made in response to "the deluge of comments" he had received both from governments and from other scientists after the final draft had been circulated.

The concluding summary was removed for reasons of consistency, says Santer, as all other chapters in the report contain just one executive summary. References to uncertainties in climate modelling were not removed, he adds. "The executive summary and four and a half pages of Chapter eight are specifically devoted to the discussion of uncertainties in estimates of natural climate variability and the expected 'signal' due to human activities."

Santer describes as a "supreme irony" of the criticism the fact that he "fought very hard during drafting sessions to include sections on signal and noise uncertainties". He points out that many of his co-authors advocated removing these sections, on the grounds that the issues were partially covered in other chapters.

Sir John Houghton, co-chair of the IPCC's science working group, says the GCC's allegations are "scurrilous" and "have absolutely no basis in fact". The IPCC's rules of procedure, he contends, allow authors to make late modifications to the text.

The rules, he adds, state that documents are not "to be approved in detail" by working groups. This is designed to allow the modification of text that does not meet the criteria that an IPCC report must be "comprehensive, objective and balanced". Leaving the chapter unmodified would have breached IPCC procedures, says Houghton.

Houghton points out that many of the revisions were prompted by what he describes as "political interference" from the GCC. "They were openly lobbying Kuwait and Saudi Arabia in Madrid to try and weaken the science [in the report]. This was resisted by the IPCC, and we have now ended up with a document that is scientifically much better."

Santer says he is angry that the GCC "is making a big stink of the whole affair", especially because the report's key phrase — "taken together, these results point towards a human influence on global climate" — was approved by all 100 participating governments and is contained in both the draft and the published IPCC report.

Santer adds: "I am really troubled by what is going on. This appears to be a skilful campaign to discredit the IPCC, me and my reputation as a scientist." **Ehsan Masood**

South African museums under review

Cape Town. A draft white paper (policy document) on arts and culture released last week by Ben Ngubane, South Africa's Minister of Arts, Culture, Science and Technology, has done little to dispel uncertainty about the future of the country's natural history museums.

The report acknowledges that the national status of the 18 museums — which are currently funded by subsidies from the department, rather than by provincial or local authorities — needs to be reviewed. Five of them have major natural history collections, as do at least five provincial museums and one municipal one.

The latter, the Durban Museum of Natural Science, is the country's most popular museum. Its director, Brett Hendey, points out that if the review is intended to lead to an eventual across-the-board rationalization, it should include other institutions that can claim national status.

Themba Wakashe, chief director responsible for museums in the department, says that the review, which has already started, will not be confined to national museums, and may be completed in time for its conclu-

sions to be incorporated into the final draft of the white paper. Museums stripped of their national status could fall under provincial or local authorities — or possibly a combination of both — he adds.

The paper also proposes that a National Heritage Council be established. It would be responsible for allocating all funding that is currently distributed by the department. This would include distributing "operational" (salary and maintenance) funds between the various national collections, and "programme" funds, for which all museums can bid, irrespective of their status.

The department will meet leading officials of the Association of Directors of National Collections (ADNC) this week to discuss how to restructure the museums, in the hope that they may be able to formulate proposals to be included in the final version of the white paper. But it is not clear how museums that do not at present enjoy national status are to be involved in this process, if at all. Nor have provincial or municipal museums been included in a survey on rationalization. **Michael Cherry**

US urged to keep labs open to foreigners

Washington. A new report from the US National Academy of Engineering has strongly recommended that the United States should keep its doors open to foreign participation in US research and development (R&D), both public and private.

"In the absence of clear threats to national security, Congress should avoid legislating restrictions on foreign participation in privately funded US R&D," says the report, which was prepared by a committee headed by Alan Schriesheim, director of the Department of Energy's Argonne National Laboratory in Illinois, and released last week. "Congress should strike reciprocity requirements from existing laws governing federal R&D spending, and exclude them from future R&D legislation."

Panel members and outside observers acknowledge that Congress is unlikely to rush to amend existing laws. But the committee wanted to send a strong message that restrictive policies do more harm than good, says Schriesheim. "I hope the report will at a minimum prevent legislation that would close doors."

The committee began work on its report nearly three years ago, when the issue of US 'competitiveness' was more politically charged than it is at present. At the time, anecdotal stories in the media suggested

that US companies were being bought up or outmanoeuvred by foreign competitors.

Legislation enacted by Congress in the early 1990s reflected this anxiety by building in reciprocity agreements to government-funded R&D programmes. This meant that foreign companies that wanted to participate in a Co-operative Research and Development Agreement (CRADA) with the Department of Energy, or in the Commerce Department's Advanced Technology Program (ATP), had first to show that their home governments provided US companies with similar access to R&D programmes.

Meanwhile, foreign investment in privately financed US R&D has been growing. The share of total domestic R&D spending by US affiliates of foreign-owned companies rose from 9.3 per cent in 1982 to 15.5 per cent in 1993, with the pharmaceutical and industrial chemical sectors recording the heaviest foreign investment. The US government has less influence in this area, except where national security is concerned. But the committee argues that this growth, as well as the proliferation of transnational corporate alliances, are, on balance, "positive trends".

The report's enthusiasm for open-door policies partly reflects a change in the political climate, says Schriesheim. The United

States is no longer seen as lagging behind foreign competitors. Global economics and the creation of multinational ventures are more firmly entrenched than ever, and even individual states are eagerly courting foreign investment. As the political atmosphere shifted, panel members who initially had advocated more restrictive stances softened. "The sentiment became more laissez-faire," says Schriesheim.

In fact, the academy report has appeared after the argument has largely ended. No new restrictive legislation has been proposed recently in Congress, and discussion of programmes such as ATP have been more concerned with survival, in a Republican-controlled Congress that is hostile to "industrial policy".

Although the committee says that it is "convinced that the strength of the US basic research system is closely linked to its openness", it claims there is still a need to deal with unfair practices in the conduct of international R&D. It calls on the government to monitor and report on US access to government-sponsored R&D in other countries, and to use trade agreements and other tools — rather than what Schriesheim calls the "blunt instrument" of restrictive policies — to ensure reciprocal access.

Tony Reichhardt

Australian industry joins scientists' protests at geology cuts

Sydney. Earth scientists in Australia and the powerful mining industry have joined forces to oppose severe cuts being proposed for the Australian Geological Survey Organisation (AGSO) in the austerity drive promised by the new conservative Coalition government.

The threat to AGSO broadens the sources of anxiety felt in the science community and technology-based industries, as the agency comes under a third ministry — that of Primary Industries and Energy — in addition to two that are already placing pressure on scientists, namely Science and Technology, and Education.

Internal documents reveal that the government is planning 'savings' that would be achieved by slashing AGSO's annual budget by 30 per cent over the next year, from A\$55.5 million (US\$43 million) to A\$39 million. This would mean sacking 130 staff — 25 per cent of the total — and axing entire survey and research programmes.

Chris Powell, head of geology and geophysics at the University of Western Australia, and president of the Australian Geoscience Council, a federation that represents 7,000 members of nine professional societies, described the proposals as "short-sighted and a bloody disaster".

Programmes or projects marked for severe shrinkage or termination include

climate change, the Cooperative Research Centres for Antarctic and Southern Ocean, and for Mineral Exploration Technologies, Antarctic geology, coastal zone studies, national geoscience information system, geohazards, palaeomagnetism and timescales.



Powell says most valuable work undertaken by the 50-year-old AGSO, previously known as the Bureau of Mineral Resources (BMR), has been pre-competitive research on mineral and petroleum deposits needed to open up Australia's vast mining potential.

One proposal would be either to terminate the work by the research vessel *Rig Seis-*

mic, or to operate it commercially in competition with private contractors — a prospect that concerns the petroleum industry, which argues that it has already paid for the vehicle once through taxes, as well as scientists, who need unrestricted open access to publicly funded surveys.

Rig Seismic is the only vessel sailing from Australian ports carrying out marine surveys to support exploration and management plans for Australian territorial waters, which have recently been expanded.

"If the cuts go ahead, the government will be squeezing out all the medium-sized companies that cannot afford to do both general surveys and targeted explorations for economic deposits," Powell claims. "The old BMR was responsible, almost single-handedly, for opening up the North of Australia and generating enormous revenues."

Staff of AGSO are said to have been "stunned" by the extent of the proposed cuts, which are proportionately three times higher than the 10 per cent target for the whole department within which AGSO operates. AGSO executives are also said to be nervous about the future of a new building worth A\$85 million. The leading industry body, the Australian Minerals Council, has already called on the government to moderate the cuts.

Peter Pockley

Holocaust homework question prompts protests in France

Paris. A French schoolteacher was suspended last week for giving her pupils as homework the task of calculating the concentration of carbon monoxide (CO) needed to gas Jews during the Second World War.

The outline of the homework question was that "Hitler killed Jews by locking them in lorries where the exhaust pipe had been connected to the interior". It went on to give the volume of the lorry, the fatal concentration of CO reached, and then asked pupils to calculate the amount of CO needed to be produced by the engine, given that "people took 20 minutes on average to die".

The incident has caused widespread protests in France. The teacher, who works at a high school near Paris, has defended the exercise as an attempt to remind the children of the Holocaust. But critics argue that her action was naive and insensitive, and could only contribute to a banalization of this tragic period of history. □

Nuclear test ban edges closer

London. Prospects for a comprehensive nuclear test ban treaty — currently being negotiated at the disarmament conference in Geneva — edged closer last week, when China unexpectedly announced that it was prepared to accept a temporary ban on 'peaceful' nuclear explosions, carried out for research rather than military purposes.

But Sha Zukang, China's ambassador to the disarmament conference, said the matter would need to be assessed at the disarmament review conference in 10 years' time before China signed any ban on all nuclear explosions, including those for military purposes. The ambassador added that China would not automati-

cally accept on-site inspections — Beijing wants inspections to be approved by two-thirds of the executive council — and would be conducting one more nuclear test before its ban came into force. □

Computer partners agree terms

San Diego. The two partners operating the San Diego Supercomputer Center in California have patched up a dispute about a new grant from the US National Science Foundation (NSF) (see *Nature* **381**, 104; 1996), and have agreed to submit a joint proposal for NSF funds.

Officials announced on 3 June that the University of California, San Diego, will lead the new proposal, while General Atomics, a private company that is the current manager, will become the major subcontractor. Sidney Karin, who has been the director of the centre since the 1985 partnership began, will direct the combined San Diego proposal for the NSF grants. The winners will be announced in the autumn. □

Hunting less fatal than pollution?

London. Whales are at greater risk from ozone depletion and atmospheric pollution than from commercial whaling, according to Peter Bridgewater, president of the International Whaling Commission (IWC). His comments came as the Environmental Investigation Agency, a conservation group, calculated that more than 1,500 striped dolphins have died in the Mediterranean in recent months from a virus linked to elevated concentrations of a pollutant. In addition, the group estimated that 750 bottle-nosed dolphins died in the Gulf of Mexico from a viral infection, toxic algae and pollution.

The debate about commercial whaling is likely to resume when whaling nations meet in Aberdeen, Scotland, for the forty-eighth annual IWC meeting from 24 to 28 June. Norway, Japan and Iceland

To **save an hour**
each time you
purify histidine-
tagged proteins,
put it on the
tip of a syringe



Motorola versus radioastronomy

SIR — I should like to add to Alison Abbott's News item, "Mobile telephones ring out changes for radioastronomy frequencies" (*Nature* **380**, 569; 1996), in which she reports on the technical conflict between Motorola's proposed Iridium global mobile telephone system and the radioastronomy community about the jamming Iridium is likely to cause in the 1,610.6–1,613.8 MHz radioastronomy band.

Motorola, in designing the Iridium system, chose to use the upper part of the 1,613.8–1,626.5 MHz band for the downlinks from their low-Earth-orbit satellites to their individual hand-held 'subscriber' mobile telephones. They are entitled to do this because the band is allocated by the International Telecommunication Union (ITU) to the Mobile-Satellite Service (MSS). But the allocation for use in the space-to-Earth direction is termed 'secondary', which means the MSS must accept interference from any service having a 'primary' allocation in the same band and it must itself, of course, not cause interference to a primary service.

Unfortunately for Motorola, the Radio Astronomy Service (RAS) has a primary allocation in the neighbouring band of 1,610.6–1,613.8 MHz, which embraces one of the spectral lines of the OH radical, and radiotelescopes are peculiarly susceptible to interference from transmissions from satellites.

To communicate directly with hand-held telephones, the Iridium satellites need to transmit an amount of power which is sufficiently significant, in terms of what is available on a satellite, for it to be necessary to use high-efficiency radio-frequency power amplifiers. But it is an unfortunate truth that all high-efficiency power amplifiers are nonlinear. This would not matter if the signal to be transmitted were of constant amplitude or in the nature of pulses, only on or off. But such an amplifier inevitably distorts a signal that varies in amplitude. Now the system architecture chosen by Motorola requires the amplifiers to handle simultaneously several independent radio-frequency signals aimed at different subscribers. Even though each signal on its own would not suffer distortion, the superposition of several, each on its own 'carrier frequency', creates a composite signal that varies in amplitude. This inevitably leads to 'intermodulation distortion' — spurious frequency components are generated both inside and outside the band occupied by the intended signals.

This would still not matter if the satellites each had only one power amplifier, as it would then be possible for it to be followed by a suitable output filter to remove

at least the 'out-of-band' spurious frequency components. But Motorola has further chosen to use 'phased arrays' on the satellites — no less than seven arrays per satellite — and each array is composed of a large number of individual amplifiers. To equip them all with adequate output filters is difficult and would add, perhaps prohibitively, to the weight of the satellites.

The power in the intermodulation products falling in the RAS band would depend on the amount of telephone traffic being carried by Iridium and is hard to estimate with any accuracy. Typically it might be a thousand times above the level acceptable to the RAS and it therefore would amount to total jamming of the band from the point of view of astronomical research.

Motorola should have recognized at an early stage of its system design that it would be impossible to build phased arrays of power amplifiers that meet the requirements of the Radio Regulations of the ITU; these state clearly that the RAS is primary and the MSS downlink is secondary, and a special blanket footnote (FN733E) specifies that "[h]armful interference shall not be caused to stations of the radio astronomy service using the band 1,610.6–1,613.8 MHz by stations of the radiodetermination-satellite and mobile-satellite services".

Motorola has stubbornly persisted with its defective system architecture. Instead of going back to the drawing-board and designing a system to higher engineering standards, it has become intent on forcing the radioastronomy community to accept curtailment of its ability to make observations in this band.

It is not clear how this conflict will end, but if Iridium goes ahead as planned it will be piracy in the radio spectrum, it will make a mockery of the ITU and its Radio Regulations, and it will be a disaster for radioastronomy.

John Ponsonby

*Onsala Space Observatory,
S-439 92 Onsala, Sweden*

Citation policy on sequences

SIR — Although editorial boards of scientific journals strive to communicate as objectively and as rigorously as possible, there is an important issue that has not been dealt with by these boards. A survey of the 'Instructions to Authors' from a variety of international journals, including *Nature* and *Science*, indicates no clear policy on the citation of sequence information. The policy may dramatically influence the measurement of both the citation index and

journal impact factors (JIFs), given the proliferation of complete-genome sequencing projects.

When citing sequence information generated by other laboratories, an author has two seemingly equivalent options; accession numbers may be quoted, or the manuscript in which the sequence was reported may be cited. As both sources are increasingly cross-referenced, the former citation may be preferred by some authors; the data will be more immediately accessible, while the original sequence data decreasingly appears in the manuscript. But the two options have drastically different effects on the sequence generator's citation index, as well as the JIF of the journal in which the original article appeared. Although it can be argued that the sequence itself is not particularly scientifically meritorious, the same criticism could easily be levelled at method papers that have nevertheless enjoyed huge citation indices; scientists simply learn to distinguish between a much-cited method paper and a much-cited paper that makes central advances in a particular field. The lack of a formal policy means that citation indices and JIFs are fuzzier measurements than are presently appreciated. The simplest policy solution is that sequences should always be cited by both publication and accession number.

A related issue concerns the assignment of priority to a particular piece of sequence data. For example, the 16S ribosomal RNA gene from the honeybee mitochondrion was originally sequenced by Vlasak *et al.*¹. Has the appropriate citation been superseded by the publication of the complete sequence of the honeybee mitochondrial genome², or do the original authors have priority? The distinction is an important one for editorial boards to consider, given the increase in complete-genome sequencing projects in recent years. It seems clear that history should have precedent, unless the original data are found to be flawed. But until editorial boards clarify this issue, citations will continue to be misappropriated.

Mark Downton

*Department of Crop Protection,
Waite Campus,
Adelaide University,
Glen Osmond, 5064, Australia
and*

*Department of Biology,
Wollongong University,
Wollongong 2522, Australia*

1. Vlasak, J., Burgschwaiger, S. & Kreil, G. *Nucleic Acids Res.* **15**, 2388 (1987).

2. Crozier, R. H. & Crozier, Y. C. *Genetics* **133**, 97–117 (1993).

corres@nature.com

Letters submitted for Correspondence should be typed, double-spaced, on one side of the paper only, or e-mailed (with full mailing address) to corres@nature.com

The ozone layer: the road not taken

Michael Prather, Pauline Midgley, F. Sherwood Rowland and Richard Stolarski

In evaluating the cost of phasing out chlorofluorocarbons, it pays to look not only at current damage to the ozone layer but also at what might have happened had our use of these chemicals continued unimpeded.

IMAGINE the Antarctic ozone hole bursting upon a scientific community in which concern about depletion of stratospheric ozone had faded out after the debate over supersonic aircraft in the early 1970s. What would today's atmosphere be like and what future might we expect? Does analysis of the road not taken help towards solving current or future global problems?

Estimates, with considerable uncertainty attached, of the costs of the global phase-out of chlorofluorocarbons (CFCs) and other man-made chlorocarbons are now appearing. In this Commentary we analyse the benefit side — where estimates are also uncertain — of the comparison of what might have happened had global use of these ozone-depleting chemicals continued unimpeded.

The threat to the global ozone layer posed by CFCs and related halocarbons has been dispelled because, since the early 1970s, the global community has followed a path of scientific understanding, public awareness, environmental activism and boycotts, national regulations, industry studies of CFC substitutes and, finally, an international agreement — the 1987 Montreal Protocol on Substances that Deplete the Ozone Layer. Without this sequence of events, which culminated first in restricted use of CFCs and later in a complete phase-out of these chemicals, ozone depletion would be worse than it is today and the global atmosphere would have been committed to a very different future over most of the next century.

Here we compare the path we are now following — the amended Montreal Protocol and its implied future — with the most likely path of CFC growth and ozone depletion expected had not the CFC threat to the ozone layer been identified in 1974. In particular, what if CFC use had followed a free-market growth into all sectors and countries until the observation of stratospheric ozone depletion itself were the first warning? What if the ozone hole had been discovered in 1985 (as indeed it was) but no prior suspicion had focused on chlorine from CFCs? When can we now expect to see recovery of the ozone layer as the provisions of the protocol show their effects in the global atmosphere? How much stratospheric chlorine and associated

ozone depletion would the global atmosphere have been committed to in a likely alternative scenario?

Early stratospheric science

Consider the state of stratospheric science in 1974. P. J. Crutzen had demonstrated in 1970 the importance of nitro-

longer envisaged, and stratospheric chlorine from projected Space Shuttle launches was predicted to have a negligible effect. Stratospheric science in the United States had received a great boost from the Climatic Impact Assessment Program, but no intense follow-on science was

High returns: early balloon experiments like this allowed researchers to point quickly to CFCs as the cause of the Antarctic ozone hole on its discovery in 1985.

gen oxides from naturally occurring N_2O in the ozone balance¹, and the potential for technology to deplete stratospheric ozone on a global scale was raised by H. S. Johnston² in an assessment of the impact of a proposed fleet of supersonic aircraft (SSTs). The exhaust from these aircraft injects nitrogen oxides and water vapour directly into the stratosphere in proportion to the size and activity of the global SST fleets. The economic outlook for SSTs failed before environmental issues were fully resolved, and only a few Concorde and no 'Boeings' were built. In the late phases of the US Climatic Impact Assessment Program, initiated in 1971 to study the stratospheric problems surrounding SSTs, chlorine-catalysed loss of stratospheric ozone was discovered³ and the impact of chlorine emissions from the solid rocket motors of NASA's Space Shuttle (and also from volcanoes) was evaluated⁴.

By early 1974, concerns about threats to the stratospheric ozone layer had diminished: large fleets of SSTs were no

needed with the demise of the large-scale SST programmes. If there had been no further focus for concern about the ozone layer, stratospheric science would have continued to develop, but much more slowly through individual research efforts funded by several agencies.

Production of CFCs

For at least two decades extending into the early 1970s, the chemical industry had expanded the global markets for CFCs, continually developing new applications. Technical journals showed no abatement in proposals for new potential markets. The applications took advantage of the inertness and volatility of the CFCs; none involved their chemical alteration or permanent confinement (that is, all were intended to be eventually released into the atmosphere). The global production of CFCs showed rapid free-market growth: in 1974, the combined annual production of CFC-11 (CCl_3F) and CFC-12 (CCl_2F_2) exceeded 0.8 megatonnes and had been growing for two

decades at a rate of about 10 per cent a year — a doubling time of about 7 years. As much as three-quarters of CFC use was in rapid-turnover applications, especially as propellants for aerosol spray cans; the rest was primarily in refrigeration, air conditioning and foam blowing; and the market for CFCs as cleaning solvents had just opened up.

Although they are certainly not precise, we can adopt typical business projections for growth (Gompertz curves) leading eventually to a saturated market. Indeed, these predictions fit the growth in CFC-11 and CFC-12 very well up to 1974. The 1975 decrease in CFC production can be ascribed in part to the global recession, but the continued depression of production compared to a free market was clearly driven by the environmental concerns described below. The economics of using existing CFC industrial plants and creating new ones are very different when the product is under an environmental cloud, and plant expansion for CFC-11 and CFC-12 essentially came to a halt in the mid-1970s. (But production of CFC-113, which was not involved in aerosol propellant use, continued to rise rapidly until 1990 — aerosols were the initial focus of environmental concern.) In the free-market case considered here, growth in production of CFC-11 and CFC-12 is projected to continue from 1974 but with gradually slowing annual growth, for example declining to 5 per cent a year by 1990. Similar free-market scenarios for the other two main ozone-depleting halocarbons, methyl chloroform (MCF; CH_3CCl_3) and CFC-113 ($\text{CCl}_3\text{FCClF}_2$), were actually followed until 1990. The top right panel of the figure shows the emissions of CFC-11, -12 and -113 expected under free-market assumptions up to 2002, and compares them with those derived from expected emissions based on reported production and the amended Montreal Protocol (labelled 'Copenhagen '92'). The choice of 2002 for the end of the free market is based on considerations of scientific developments and the pace of regulatory response to them (see below).

Atmospheric build-up

The atmospheric build-up of CFCs and MCF can be reliably computed with knowledge of their emissions. Most uses of MCF and CFC-113 lead to rapid release of these compounds into the atmosphere. Three-quarters of pre-1974 use of CFC-11 and CFC-12 was also in rapid-release products such as aerosol spray cans. So in projecting the free-market case from 1974, we ignore the banking of unreleased CFCs and assume that most production would be released into the atmosphere within a year. The tropospheric chlorine loading⁵ is the sum of the concentrations of chlorine atoms delivered to the stratosphere by all halocarbons, including the natural compound methyl chloride (CH_3Cl), and so is a

benchmark of chlorine-driven ozone depletion. In the bottom right panel of the figure, the projected chlorine loading in the free-market case is compared with the actual history of atmospheric halocarbons up to 1995, appended with the projected chlorine loading under the amended Montreal Protocol ('Copenhagen '92').

The CFC threat

In 1974, the CFC threat to the ozone layer was raised by M. J. Molina and F. S. Rowland⁶. In brief, CFCs are long-lived gases that accumulate in the atmosphere, break down only in the stratosphere and release their chlorine there. Chain reactions initiated by chlorine atoms and involving ClO then act to deplete the ozone layer. In the late 1970s, CFC markets for spray can propellants were shut down in the United States, Canada, Sweden and Norway because of environmental pressure or legislation, or a combination of both. The Clean Air Act of 1977 authorized NASA to study the ozone layer and report biennially to Congress, and thus was born the Upper Atmosphere Research Program. As a result of this highly focused research programme and related efforts funded by other US federal agencies and in other countries around the world, many scientists embarked on careers to study the ozone layer, stratospheric dynamics, atmospheric chlorine and CFCs.

A better understanding of the stratospheric ozone layer arose through: (1) extensive laboratory work on the chemical reactions controlling ozone, the photochemical destruction of CFCs, and reactions of the applicable chlorine-containing chemical species; (2) instrument development and measurement campaigns directed at measuring CFCs, stratospheric chlorine and particularly ClO (ref. 7); (3) study of the data from an extensive set of ground-based ozone instruments (Dobson spectrophotometers) put in place during the 1957 International Geophysical Year and the ensuing decade; (4) the launch of TOMS/SBUV satellite instruments to measure ozone globally; and (5) the development of atmospheric chemistry models.

In the late 1970s, there were several country assessments of the potential impact of human activity on the ozone layer. In 1985, these assessments became international, and the resulting three-volume 'blue books'⁸ summarize a decade or more of intensive research on understanding and predicting ozone depletion. With incomplete knowledge of stratospheric chemistry in the early 1980s, ozone depletions of several per cent were predicted by the end of the next century if CFC production remained at 1974 levels (eventual chlorine loading of about 8 parts per billion (p.p.b.)). But when growth in CFC production resumed as in

the early 1980s, some calculations of this long-term ozone depletion exceeded 10 per cent, and control of the growth in CFC production became a more public issue.

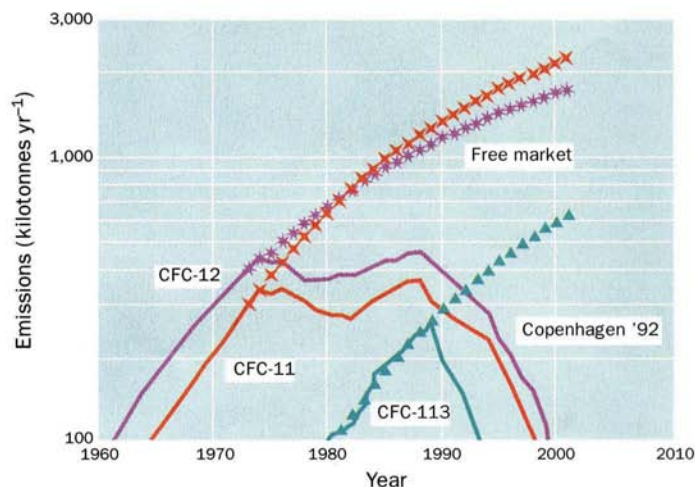
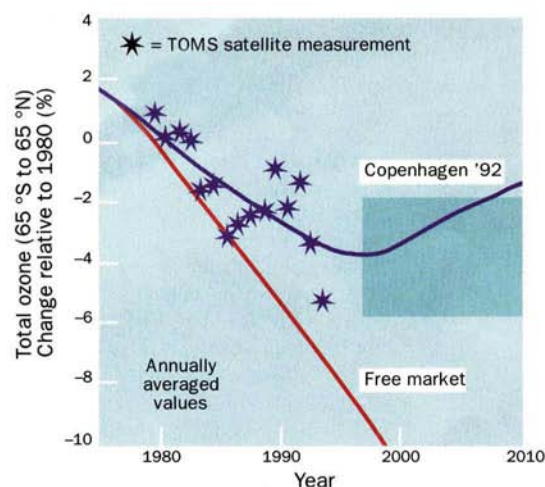
The Antarctic ozone hole

The Antarctic ozone hole came to the sudden attention of the scientific community in 1985. J. C. Farman, B. G. Gardiner and J. D. Shanklin⁹ documented the rapid springtime decrease in Antarctic ozone over their British Antarctic Survey (BAS) station at Halley Bay. Noting previous arguments for massive ozone loss at high chlorine levels, they suggested that the observed depletion was due to the rising stratospheric chlorine levels accompanying CFC growth. The BAS observations were rapidly confirmed by NASA satellite data and shown to extend over the entire Antarctic polar region¹⁰. The ozone hole has recurred every year since and has become more pronounced. None of the early 1980s' atmospheric models, all of which considered only homogeneous gas-phase chemistry, could explain the amount of ozone lost when sunlight returned to Antarctica each spring. In 1986, a novel mechanism was put forward in which chemical reactions on polar stratospheric clouds converted inactive chlorine reservoirs such as HCl into reactive ozone-destroying forms such as ClO (ref. 11). But this mechanism alone was not sufficient to explain the rapid ozone loss. Subsequent laboratory work¹² identified a new chain-reaction mechanism whereby ozone could be quickly and nearly completely removed by ClO at high concentrations. The patterns of ozone depletion expected to result from these combined mechanisms were later measured inside the Antarctic ozone hole¹³.

During this period there was open scientific debate about the cause of the ozone hole: some scientists argued for mechanisms driven by meteorological changes, solar-cycle effects or other chemical processes. Continued measurements and modelling soon eliminated these other possibilities. So fairly soon after the discovery of the Antarctic ozone hole, the scientific community had abundant evidence that the primary cause of the depletion was the increase in CFCs and related halocarbons.

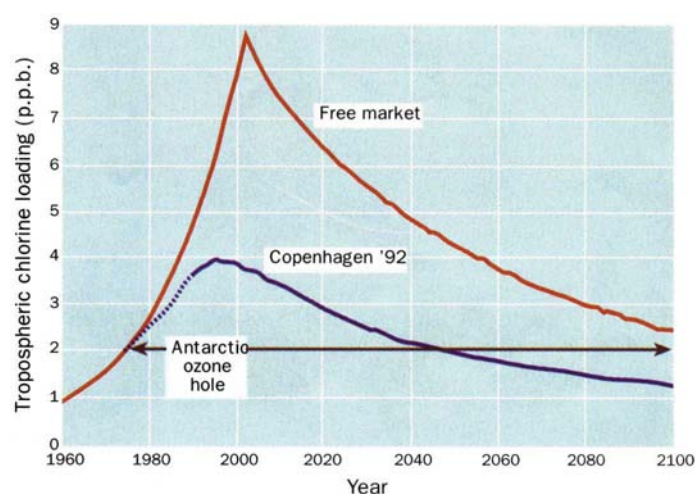
Protocols and phase-outs

The Montreal Protocol was being negotiated at the time of the discovery and explanation of the Antarctic ozone hole. Its ratification and subsequently stronger amendments and adjustments were made possible only through the convincing scientific evidence of ozone depletion occurring globally, as well as over Antarctica. The complete phase-out of CFC and MCF production (as well as some other



Top right, emissions (kilotonnes yr^{-1}) of chlorofluorocarbon (CFC) -11 (orange line, crosses), -12 (pink line, asterisks) and -113 (turquoise line, triangles). The lines refer to the Copenhagen '92 scenario based on historical production and projections following the amended Montreal Protocol. The symbols extend each of the CFCs emissions with 'free-market' projections from 1974 to 2002. Bottom right, tropospheric chlorine loading (parts per billion (p.p.b.)); the sum of chlorine atoms in CFC and related halocarbons) following the emissions from the Copenhagen '92 and free-market scenarios in the top right panel. The chlorine level above which we expect the annual formation of the Antarctic ozone hole is shown by the horizontal arrows at about 2 p.p.b. of chlorine. Top left, percentage global mean ozone depletion relative to 1980, averaged annually and over the latitudes 65°S to 65°N, as measured by the TOMS/SBUV satellite (asterisks) and as predicted by the Goddard 2D model (C. H. Jackman, personal communication). The recovery of ozone expected under the amended Montreal Protocol (Copenhagen '92) is shown as a purple line, and that for the free-market chlorine levels is shown by the red line.

Based on the observed year-to-year variability in the total ozone, a band (darkly shaded area) has been applied to the projected ozone-recovery period beginning in 1997 to determine when it might fall outside this range and be readily detected.



halocarbons) was agreed in London (1990) and was brought forward by four years in Copenhagen (1992), with the result that CFC production in developed countries was ended by 1 January 1996. A few months more than 8 years elapsed from the initial Montreal Protocol agreement to the complete phase-out of CFCs. Once agreement had been reached on control or elimination of CFCs, market incentives accelerated the phase-out; and the chemical industry attained the established goals ahead of time by more rapidly switching over to alternative compounds. The period 1987–1995 was required to convince all parties of the scientific and environmental needs and to develop and switch to alternative compounds. A best approximation to the historical and projected emissions of CFCs and related industrial halocarbons is a bottom-line case, here designated 'Copenhagen '92'. Emissions of CFCs are shown in the top right panel of the figure, and the resulting tropospheric chlorine loading is shown in the bottom right panel.

Without early warning

Without early warning of the causal relationship between CFCs and stratospheric ozone depletion by Molina and Rowland in 1974 — or soon after by someone else — stratospheric science would have evolved differently. Without a pressing concern such as CFCs, the overall pace of investigations of stratospheric chemistry would probably have slowed considerably over the decade 1975–1985. In 1971, when the first CFC paper appeared¹⁴, no measurements had yet been made of any chlorine-containing compound in the stratosphere, nor were the tropospheric data yet calibrated well enough for the observed growth in the tropospheric concentration of these chemicals to stimulate interest. (Even when people did become interested in their tropospheric growth rates, such calibration waited several more years.) Without direct enquiry into the atmospheric fate of CFCs, the likely triggering event would have been the observation of springtime Antarctic ozone loss by the BAS observational system. Yet without

the preceding 11 years of scientific publications on the stratospheric effects of chlorine from CFCs, Farman and colleagues almost certainly would not have pointed to chlorine or CFCs as the cause of the ozone hole, and their observations might never have made the headlines — at least not until the ozone hole as an Antarctic-wide phenomenon had been corroborated by the global maps of ozone from the TOMS satellite.

How long would it have taken to establish CFCs as the source of stratospheric chlorine, to measure ClO in the stratosphere (and inside the ozone hole), to use these observations with laboratory measurements to define the mechanism of ozone loss, and to prove beyond reasonable doubt that CFCs were the cause of the ozone hole? Clearly, a substantial time lag would have transpired while the developments in scientific knowledge we actually made during our decade 1974–1985 were discovered for the first time along this alternative path. The instruments, and even the techniques, used for measuring chlorine compounds

in the stratosphere in 1985 were simply not in existence in 1974, nor was there any particular motivation to develop them. The successful measurement of ClO from the ER-2 aircraft in 1987 had its origins in the balloon-borne instruments first proposed and funded in 1975.

Without the early warning, there would also have been no demands to phase-out CFC use in aerosols, and CFC production would have been expected to follow the free-market scenario. The elapsed time from early warning to the Montreal Protocol was 13 years, and to the full regulatory ban in the developed world it was more than 21 years. If the first alert had indeed been the BAS observations of Antarctic ozone loss in mid-1985, then one could argue that the full regulatory ban might have been set for 1 January 2007 — that is, 21 years later. Allowing for some acceleration in our learning, assume that the 1975–1985 knowledge was gained in 8 years, that is by 1993. Optimistically, it would still take a further 2 years (1995) to adopt a protocol and another 7 years (2002) to shut down the CFC industry. This second period would be critical: we would have to continue to build convincing evidence that CFCs were responsible for the ozone loss and also find industrial alternatives that could fulfil their uses. Consider the repercussions to ozone, and even to climate forcing, of allowing CFC use to grow until 2002 or even longer, say to 2007.

Environmental consequences

If we had ignored the scientific evidence, or not had it at all, ozone depletion would be dramatically worse than that which we are controlling today. Compare the levels of atmospheric chlorine expected under the Copenhagen '92 scenario with the alternative free-market scenario that shuts down the CFC industry only after 2002. The chlorine loading in the latter situation would approach 9 p.p.b. in the first decade of the twenty-first century instead of a maximum of just under 4 p.p.b. in the 1990s. The predicted globally averaged ozone depletion relative to 1980, shown in the top left panel of the figure, would exceed –10 per cent soon after 2000, rather than just reaching –4

per cent in 1997. The cumulative impact would be even greater, because the larger ozone depletion in the free-market case would also recover more slowly, lasting well into the twenty-first century. At these levels of chlorine loading, our predictions are still scientifically sound but are more uncertain because such stratospheric conditions lie outside our range of experience.

For example, at 9 p.p.b. of chlorine, the Arctic lower stratosphere might become dominated by chlorine instead of nitrogen oxides (as now happens over Antarctica owing to denitrification), and during very cold Arctic winters the ozone loss would be more difficult to predict. If CFCs had followed free-market growth until 2002, the Antarctic ozone hole would be a permanent fixture throughout the twenty-first century, instead of disappearing by 2050 as predicted in the Copenhagen '92 scenario. Additionally, CFCs are important greenhouse gases, and for the period 1985–2000 their increases would have added more radiative forcing to the climate than that due to typical increases in carbon dioxide (+1.5 parts per million a year). (By contrast, over the same period in the Copenhagen '92 scenario, CFC increases contribute only about 40 per cent of the additional radiative forcing.) Much of this additional greenhouse forcing would, like the ozone hole, have remained throughout the next century. So the future not taken is much worse than

Given the cold shoulder: over 21 years elapsed from the early warning about CFCs to the shutdown of the industry in them.

André Maslennikov/Still Pictures

that expected in the Copenhagen '92 scenario, and has no easy repair.

The good news is that we, as a global community, are committed to following the amended Montreal Protocol and the phase-out of CFCs. Effectively, we are now in the second phase of our grand experiment, that of reversing our atmosphere's CFC content. Ozone depletion is expected to reverse and recover measurably in the first decades of the next century, as shown in the top left panel of the figure. This prediction is part of a causal chain whose verification is an important part of the overall scientific assessment process: production and emission of CFCs have fallen greatly; the tropospheric chlorine content has now peaked; we expect the peak in stratospheric chlorine to follow a few years later; and stratospheric ozone should recover soon after. Because of the year-to-year variability in global annual mean ozone (see top left panel of the figure), unambiguous detection of the recovery will take a decade or more. If we do not make careful measurements during that time and do not make the effort to understand and characterize this data, then we will not be able to demonstrate the recovery of the ozone as soon as possible and close the chapter on the CFC threat to the ozone layer. □

Michael Prather (Department of Earth System Science) and F. Sherwood Rowland (Department of Chemistry) are at the University of California, Irvine, California 92717-3100, USA. Pauline Midgley is an AFEAS consultant at Ludwigstrasse 49, D-70771 Leinfelden Emsingen, Germany. Richard Stolarski is at the NASA Goddard Space Flight Center, Greenbelt, Maryland 20771, USA.

ACKNOWLEDGEMENTS. This Commentary is based on a session at the December 1995 American Geophysical Union meeting and has benefited from comments from W. Brune, P. Fraser, C. Jackman, M. McFarland, S. Montzka and R. Prinn.

- Crutzen, P. J. *Q. J. Roy. Met. Soc.* **96**, 320–325 (1970).
- Johnston, H. S. *Science* **173**, 517–522 (1971).
- Stolarski, R. S. & Cicerone, R. *Can. J. Chem.* **52**, 1610–1615 (1974).
- Committee on Impact of Stratospheric Change. *Halocarbons: Effects on Stratospheric Ozone* (National Academy of Sciences, Washington DC, 1976).
- Prather, M. J. & Watson, R. T. *Nature* **334**, 729–734 (1990).
- Molina, M. J. & Rowland, F. S. *Nature* **249**, 810–814 (1974).
- Anderson, J. G., Margitan, J. J. & Stedman, D. H. *Science* **198**, 501–503 (1977).
- Global Ozone Research and Monitoring Project Report No. 16: Atmospheric Ozone 1985* (World Meteorological Organization, Geneva, 1985).
- Farman, J. C., Gardiner, B. G. & Shanklin, J. D. *Nature* **315**, 207–210 (1985).
- Stolarski, R. S. *et al. Nature* **322**, 808–811 (1986).
- Solomon, S., Garcia, R. R., Rowland, F. S. & Wuebbles, D. J. *Nature* **321**, 755–758 (1986).
- Molina, L. T. & Molina, M. J. *J. Phys. Chem.* **91**, 433–436 (1987).
- Anderson, J. G., Brune, W. H. & Proffitt, M. H. *J. Geophys. Res.* **94**, 11,465–11,479 (1989).
- Lovelock, J. E. *Nature* **230**, 379 (1971).

An old galaxy in a young Universe

Robert C. Kennicutt Jr

A mature galaxy has been discovered in an early phase of the Universe apparently too young to contain it. Is this the end of the theorists' favourite cosmology, the Einstein–de Sitter model?

DEEP images from the Hubble Space Telescope and other instruments have provided unprecedented glimpses of the Universe at high redshifts, when it was a small fraction of its present age. Not surprisingly, most of the distant galaxies revealed in these spectacular images bear the signatures of youth, with blue colours dominated by the light of young stars, a striking variety of shapes, and pervasive evidence of interactions and mergers¹. But these same surveys reveal a handful of red, apparently old galaxies, which tell a different but equally compelling story. On page 581 of this issue², Dunlop *et al.* report the discovery of the most perplexing case to date, a galaxy at a

redshift of 1.5 with an apparent age of at least 3.5 billion years.

The problem, if conventional cosmological models are correct, is that galaxies that old and that far away simply should not be there. The observation tightens the thumbscrews on the Einstein–de Sitter cosmological model (see box), and offers evidence that at least some galaxies formed at very early epochs, within a billion years after the Big Bang.

The galaxy discovered by Dunlop and colleagues, 53W091, was first identified in a deep survey of faint radio galaxies — radio selection has proved effective in identifying distant galaxies with a range of

evolutionary properties^{3,4}. Imaging of 53W091 in the visible and infrared revealed extremely red colours, redder than a young galaxy at any plausible redshift and consistent with an evolved galaxy observed at high redshift. A more precise redshift, $z = 1.552$, came from a 5.5-hour spectrum taken with the ten-metre Keck telescope. The successful measurement of the spectrum and redshift of this object is noteworthy in itself: at 26th magnitude, this is one of the faintest galaxies ever observed spectroscopically, and probably the faintest to have its redshift determined using stellar absorption lines.

It is far from being the most distant

A meeting of Hubble constants

ONE of the original goals of the Hubble Space Telescope (HST) was to determine precisely the expansion rate of the Universe, the Hubble constant (H_0). After delays, optics problems and a succession of conflicting H_0 measurements, many have wondered whether the controversy will be settled within the lifetime of HST. Now comes news, presented at a conference last month*, that the measurements of H_0 may be converging at last.

Measuring H_0 is difficult because it requires an accurate distance scale out to at least 100 megaparsecs (300 million light years), far enough to reliably trace the cosmological expansion. There are several secondary standard-candle distance indicators, which must be calibrated against a primary scale provided by Cepheid variable stars in nearby galaxies. But the resulting scales have disagreed by up to a factor of two, mainly due to a dearth of precise Cepheid distances. HST is gradually solving this problem by providing such distances for about 25 galaxies.

Distances are now available for about half of that sample. The H_0 Key Project presented several new distances, including that of NGC1365 in the Fornax cluster (see photograph). Applying these distances to several secondary methods gave an H_0 of 68 to 77 $\text{km s}^{-1} \text{Mpc}^{-1}$ (J. Mould, Mt Stromlo Obs. and W. Freedman, Carnegie Obs.). A competing group,

headed by Allan Sandage, based most of their results on a single secondary method, the apparent brightness of type Ia supernovae, and find $H_0 = 55$ to 61 $\text{km s}^{-1} \text{Mpc}^{-1}$ (G. Tammann, Basel Univ. and A. Saha, STScI). These are higher than



W. Freedman/HST/NASA

the 43 to 53 $\text{km s}^{-1} \text{Mpc}^{-1}$ values reported previously, whereas the latest Key Project results are about 10% lower than their preliminary estimate. The convergence mainly reflects the rapidly improving Cepheid distances from HST.

Refinements to other secondary methods were reported, including surface-brightness fluctuation (J. Tonry, MIT), planetary nebula luminosity function (G. Jacoby, Kitt Peak), the Tully–Fisher relation (R. Giovanelli, Cornell Univ.) and other supernova methods (R. Kirshner, Harvard Univ.), all yielding H_0 between 65 and 82 $\text{km s}^{-1} \text{Mpc}^{-1}$. Most agree well with the primary Cepheid distances for nearby objects, indicating that systematic errors — long the problem in this field — are being

eliminated. These values are still 20 to 30% higher than the 'long' scale favoured by Sandage and colleagues, but that is far from the factor-of-two disagreement that existed only 3 years ago.

As the uncertainty in H_0 decreases, attention turns to its implications for the cosmological age problem. For a Friedmann cosmological model, in which gravity is determined by matter alone, $H_0 = 70 \text{ km s}^{-1} \text{Mpc}^{-1}$ implies a cosmic age of 9 to 14 Gyr, depending on the density of the Universe. For the special-case Friedmann model where the density is just enough to close the Universe — the flat Einstein–de Sitter model favoured by inflationary cosmology — the expansion age is 9 Gyr. This is flatly inconsistent with the ages of Galactic globular clusters, estimated to be about 13–17 Gyr (P. Demarque, Yale Univ.).

How do we solve this conundrum? If the stellar ages are correct, the Einstein–de Sitter model requires a global $H_0 < 40 \text{ km s}^{-1} \text{Mpc}^{-1}$, which now seems unlikely (M. Postman, STScI). The age problem can be avoided, barely, if the density of the Universe is much lower — less than 20% of the closure density — or if most of it is in the form of vacuum energy, the 'cosmological constant' first proposed and later abandoned by Einstein. Current observations cannot discriminate between these alternatives (M. Turner, Univ. Chicago), but there is now hope that HST may finally fulfil the mission for which it was designed 20 years ago. R. C. K.

*Space Telescope Symposium on The Extragalactic Distance Scale, Baltimore, Maryland, USA, 7–10 May 1996.

galaxy known. Star forming galaxies have been observed out to $z > 3$ (ref. 5). What makes 53W091 unusual is the combination of redshift and age. Dunlop and colleagues were able to estimate the age by fitting the Keck spectrum to stellar population models, taking advantage of the age-sensitivity of the near-ultraviolet spectrum (1,900 to 3,300 Å in the galaxy's rest frame).

The dating is complicated by uncertainty in the models and by the effects of interstellar dust and chemical composition, which can mimic an age signature, but the authors were able to circumvent much of this ambiguity by directly fitting the absorption spectrum as well as the visible and infrared colours. Their best age estimate is 3.5 Gyr, with a strong lower limit of about 3 Gyr. Most systematic errors would make the age even larger. This is not the first report of an intrinsically old galaxy at high redshift^{4,6}, but it is the strongest case, in terms of the robustness of the age determination and the high redshift of the galaxy.

The cosmological conundrum that this result presents becomes apparent when one compares this result with the expected age of the Universe at $z = 1.55$. The cosmic age at a given redshift depends on the Hubble constant and the density of the Universe. For a Universe with critical density (Einstein-de Sitter model) and H_0 in the range 60 to 80 km s⁻¹ Mpc⁻¹, the Universe is only 2.0 to 2.6 Gyr old at $z = 1.55$, at least a billion years younger than 53W091. Including a time delay of 1 Gyr to form stars increases the age discrepancy to 2 to 3 Gyr, and makes 53W091 roughly twice as old as an Einstein-de Sitter Universe at that redshift.

Many readers will recall that this is not the first instance of a cosmological 'age problem'. Values of H_0 in the range 60 to 80 km s⁻¹ Mpc⁻¹ correspond to a cosmic expansion age of 8 to 11 Gyr, far less than the 13- to 17-Gyr ages of Galactic globular clusters. These new results show that a similar age problem exists in the high-redshift Universe. The presence of luminous quasars at redshifts up to $z = 4.9$ (0.6 Gyr after the Big Bang for an $H_0 = 70$ km s⁻¹ Mpc⁻¹ Einstein-de Sitter model) may impose yet another age problem⁷. All of these inconsistencies, including the new challenge from 53W091, could be explained if we live in a low-density Universe or one dominated by a cosmological constant (see box).

The apparent age of 53W091 also imposes interesting constraints on galaxy formation and evolution models. As Dunlop *et al.* point out, the galaxy must have formed at an enormously high redshift, $z > 4$, and the subsequent star formation must have been largely completed by $z = 1.55$, a span of only 1.5 to 3 Gyr, depending on the cosmology adopted.

An observation with such important consequences begs to be confirmed, and, with the dedication of the second Keck telescope last month, further results should not be long in arriving. Other distant red galaxies need to be analysed to

ensure that the great age inferred for 53W091 is not a fluke. Even stronger limits may come from infrared imaging of a class of $z > 3$ galaxies identified recently using Keck⁵. Those objects are all active star-forming galaxies, but if they possess an appreciable infrared excess from older stars they may impose even stronger constraints on the cosmic age scale, and provide further clues to the early stages of galaxy formation. □

Robert C. Kennicutt Jr is at the Steward Observatory, University of Arizona, Tucson, Arizona 85721, USA.

EARTHQUAKES

Faults without friction?

Christopher H. Scholz

HEAT-FLOW measurements near the San Andreas fault fail to detect any frictional heating^{1,2}. This is a long-standing problem in tectonics that has come to be known as the San Andreas fault heat-flow paradox. The conventional interpretation is that the shear stresses that exist on the fault during sliding must be at least an order of magnitude less than the 100 MPa or so expected from laboratory measurements of rock friction, but just how that happens is baffling. Two papers published in *Nature* this year, one by Melosh³ and the other by David Scott (on page 592 of this issue⁴) propose solutions to the puzzle.

Scott has constructed a granular model of the brittle crust, assuming *a priori* that particles may rotate freely without friction, meshing together like interlinked cog-wheels. The model successfully predicts localized shear zones and earthquake-like elastodynamic failures, all without frictional heating — the accumulated strain energy goes into elastic waves instead. The rotation of stresses across the shear zone that is often seen in nature is also reproduced. This appears to offer a possible basic mechanism for the occurrence of earthquakes, but the question is whether such a granular model is a good representation of the brittle crust, and whether such rolling behaviour can actually occur during earthquake slip.

On what scale might such rollers exist? There are large-scale rotating blocks of the order of 100 km across within transcurrent shear zones⁵ (where the slip motion is horizontally along the fault) but they certainly don't roll during earthquakes. At the other extreme there are the particles called fault gouge, the detritus produced by fault wear; but these sub-millimetre particles are many orders of magnitude smaller than the fault core thickness of around 0.1–1 km, and are unlikely to act as rollers. Nor are they observed to in the laboratory, where it is

found that shear, with an ordinary friction coefficient, is the primary mode of deformation of fault gouge^{6,7}. So it is uncertain whether there are geological objects of an appropriate scale to act as rollers during coseismic slip.

Melosh's proposed mechanism³ is dynamic weakening of the fault by acoustic fluidization. His idea is that seismic waves generated by earthquake motions are trapped within the fault core, and temporarily reduce the stresses perpendicular to the fault so that slip can occur at low shear stress. Another proposed dynamic weakening mechanism, observed in foam rubber experiments⁸, is that the earthquake propagates like a wrinkle in a carpet, with an opening displacement mode that reduces the perpendicular stress so that shear motion can occur frictionlessly. This has been found to be a permissible propagation mode on an interface between different elastic media⁹, as is likely to be found in large-displacement faults like the San Andreas.

Dynamic weakening mechanisms such as these, if they occur, cannot be the

1. Brune, J., Henyey, T. & Roy, R. *J. geophys. Res.* **74**, 3821–3827 (1969).
2. Lachenbruch, A. & Sass, J. *J. geophys. Res.* **85**, 6185–6222 (1980).
3. Melosh, H. *J. Nature* **379**, 601–606 (1996).
4. Scott, D. R. *Nature* **381**, 592–595 (1996).
5. Nicholson, C., Seeber, L., Williams, P. & Sykes, L. *J. geophys. Res.* **91**, 4891–4908 (1986).
6. Marone, C., Raleigh, C. & Scholz, C. *J. geophys. Res.* **95**, 7007–7025 (1990).
7. Beeler, M., Tullis, T., Blanpied, M. & Weeks, J. *J. geophys. Res.* **101**, 8697–8715 (1996).
8. Brune, J., Brown, S. & Johnson, P. *Tectonophysics* **218**, 59–67 (1993).
9. Andrews, D. J. & Ben-Zion, Y. *Eos* **76**, 415 (1995).
10. Zoback, M. *et al. Science* **238**, 1105–1111 (1987).
11. Zoback, M. & Healy, J. *J. geophys. Res.* **97**, 5039–5057 (1992).
12. Scholz, C. H. & Saucier, F. *J. geophys. Res.* **98**, 17867–17869 (1993).
13. Rice, J. R. in *Fault Mechanics and Transport Properties of Rocks* (eds Evans, B. & Wong, T.-F.) 475–503 (Academic, London, 1992).
14. Sleep, N. & Blanpied, M. *Nature* **359**, 687–694 (1992).
15. Scholz, C. H. *Nature* **359**, 677–678 (1992).
16. Williams, C. & Narasimhan, T. *Earth planet. Sci. Lett.* **92**, 131–143 (1989).

1. Abraham, R. G., Tanvir, N. R., Santiago, B. X., Ellis, R. S., Glazebrook, K. & van den Bergh, S. *Mon. Not. R. astr. Soc.* **279**, L47–L52 (1996).
2. Dunlop, J. *et al. Nature* **381**, 581–584 (1996).
3. McCarthy, P. J. *A. Rev. Astr. Astrophys.* **31**, 639–688 (1993).
4. Lilly, S. J. *Astrophys. J.* **333**, 161–167 (1988).
5. Steidel, C. C., Giavalisco, M., Pettini, M., Dickinson, M. & Adelberger, K. L. *Astrophys. J.* (in the press).
6. Hamilton, D. *Astrophys. J.* **297**, 371–389 (1985).
7. Turner, E. L. *Astr. J.* **101**, 5–17 (1991).

whole answer. Because they offer no mechanism for the static friction of the fault to be low, if they were the solution to this paradox we would observe huge stress drops of the order 100 MPa during an earthquake, whereas observed stress drops are only of order 10 MPa. Given this fact, the static friction of the fault must be low, and there are other observations that suggest this is true for the San Andreas fault¹⁰, but not for less developed faults, which appear to have ordinary friction coefficients^{11,12}. But if static friction on the fault is low, solution of the heat-flow paradox does not require any dynamic weakening mechanism.

The most prevalent class of models to

explain the low static strength of the fault is that the fault core contains fluid at nearly lithostatic pressures, which lowers the effective perpendicular stress across the fault and hence the frictional resistance^{13,14}. As I pointed out in an earlier News and Views article¹⁵, however, this explanation contains its own enigma. To maintain near-lithostatic fluid pressures within the fault core, there must be a dynamically impermeable seal between the fault core and the surrounding rock, and because at the seal itself the fluid pressure will exceed the compressive stress in some directions, the seal must also be impenetrable by hydrofracturing. There is no known or imagined geological

material with those properties.

In the 25-year history of the San Andreas heat-flow paradox, no solution has appeared that is plausible and does not contain a fatal flaw or lead to another enigma. As in all such cases, either we are missing something very simple or the Earth works in ways very different from the way we think it does; or perhaps there is something wrong with the paradox itself. One possibility is that fluid flow in the crust disperses the frictional heat¹⁶, so that it does not produce an anomaly. □

Christopher H. Scholz is at the Lamont-Doherty Earth Observatory, Columbia University, Palisades, New York 10964, USA.

OBITUARY

Julius Marmur (1926–96)

WITH the death of Julius Marmur on 20 May, molecular and microbial biology has lost one of its leading practitioners. In particular, he had a great influence on studies of DNA following the discovery of its structure in 1953.

Marmur was one of those many outstanding talents acquired by North American science before the Second World War. He was born in Poland, but with his family emigrated to Canada shortly thereafter; he took undergraduate and master's degrees at McGill University, Montréal, before moving to Iowa State University where he received his PhD. In the early 1950s, following hard on Oswald Avery's discovery that DNA was the genetic material, with Rollin Hotchkiss of Rockefeller University he demonstrated that pneumococcal DNA carried the information for a specific enzyme and that the coding for two enzymes could be carried on a single piece of DNA.

With a firm grounding in bacteriology, Marmur arrived in my laboratory at Harvard in 1957. This was a time when we were exhausting what could be done with the crude preparations of mammalian DNA of those early times. We had shown that DNA in solution undergoes a dramatic change in viscosity, sedimentation rate and ultraviolet absorbance when heated above a narrow temperature range, indicating the collapse or melting of a one-dimensional crystal — the double helix. Sensing that the rapid brownian motion within the single-stranded, flexible, nucleotide chain could allow the exploration of many interchain contacts, a tiny fraction of which might correspond to the original in-register pairing that would allow the double helix to reform, we searched for this phenomenon by slowly cooling the denatured form. However, little change was found in the short times we observed. It was only with the availability of bacterial DNA, which Marmur prepared with great dexterity, that we met

with success: the much lower complexity of the bacterial DNA allowed the helix to reform in reasonable times, optimally at 25 °C below the melting temperature.

With Marmur in the lead, joined by Carl Schildkraut and others, it was possible to establish that the renatured DNA regained its biological (transforming) activity, that mixtures of normal and density-labelled

more than 5,000 reprint requests.

After three years at Brandeis University in Massachusetts, during which he continued to collaborate closely with my laboratory, Marmur moved to Albert Einstein College of Medicine, New York, as professor of biochemistry and molecular genetics. Of the highlights of his three decades of work there, two stand out. With his students he brought the yeast *Saccharomyces cerevisiae* into the molecular biological research arena, developing the standard method for the isolation of its DNA and demonstrating that its mitochondrial DNA was sharply altered, and sometimes absent, in *petite* mutants.

More recently, Marmur turned to exploring carbohydrate metabolism in yeast. This work unravelled the complex *MAL6* (maltose) locus, with its linked permease and regulatory genes, and promises insights into the evolutionary relationships hidden in its alternative carbohydrate metabolic pathways.

Marmur, however, was much more than a scientific pioneer: he was a remarkably generous and responsible man, who touched deeply all those who came within his broad ambit. His constant presence in the laboratory, his teaching by example and by questioning, and his excellence in writing and editorial work, as well as his voracious appetite for the literature (often benefiting his colleagues), made him an ideal mentor. As a result, Marmur's laboratory and his wider circle of friends became the seed bed for dozens of people who matured into highly trained, motivated and imaginative researchers, and who accepted the obligations of good citizenship in the scientific community as a privilege.

Paul Doty

Paul Doty is in the Department of Biochemistry and Molecular Biology, Harvard University, Cambridge, Massachusetts 02138, USA.

IMAGE
UNAMIGABLE
FOR A COPYRIGHT
FOR REASON
REASONS

Gerald S. Cohen

DNA formed true hybrids of intermediate density, and that limited hybridization occurred between strands of related bacteriophages and bacteria. Thus, renaturation was set to become a primary means of manipulating DNA, later playing an essential role in developing recombinant DNA, DNA sequencing and amplification by the polymerase chain reaction. Moreover, with the variety of bacterial DNAs available, Marmur and others were able to show the striking dependence of melting temperature and buoyant density on the DNA composition — that is, the guanine + cytosine content.

All in all, Marmur co-authored 21 papers in this period. His own paper on the preparation of bacterial DNA elicited

An affinity for learning

David H. Margulies

CHEMISTS and physicists have been known to remark that biology — even molecular biology, which underpins most of the subject as a whole — remains a 'soft' science. The reason is that so much work measures a distant result of many complex processes. More proximal measurements can of course be technically very difficult to make, but at least in one area of immunology progress is being made. On page 616 of this

Mechanistic models to explain the developmental fate of thymic T cells are based on the phenomena of positive and negative selection, events that are dependent on the interaction of developing thymocytes with thymic APC that either rescue them from programmed cell death or promote it. These models fall into two major categories: those governed by qualitative changes in the signal delivered by the TCR

on binding one or another MHC/peptide ligand; and those determined by quantitative differences in the strength of the interaction between the TCR and different MHC/peptide ligands.

Over the past few years, strategies for producing the appropriate reagents and for measuring the appropriate parameters have been developed. Building on this work, Alam *et al.* now report the strongest evidence to date directly correlating the affinity of the interaction between a TCR and MHC/peptide ligands with the outcome of thymic selection. For the time being, quantitative models are attractive, and they promise to tell us a great deal.

Alam *et al.* have used soluble, purified preparations of the main components of the TCR-MHC/peptide interaction to explore the kinetics and affinity of binding of a well-characterized TCR to MHC/peptide complexes in which the bound

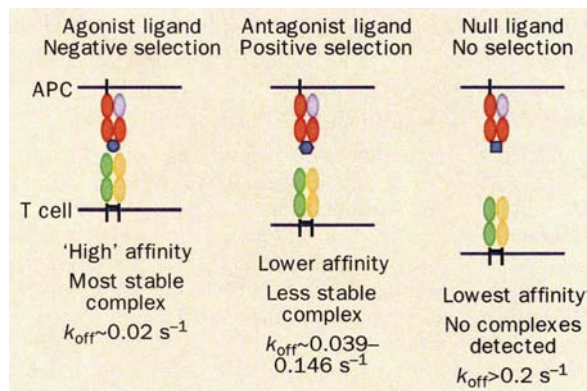
of mature MHC class I restricted cells.

Using an evanescent wave biosensor to monitor the binding of the various MHC/peptide complexes to the immobilized specific TCR, Alam *et al.* show that the interaction of the MHC/peptide complex (consisting of the wild-type ovalbumin-derived peptide bound to a purified soluble version of H-2K^b) with the TCR is weak. It is characterized by a rapid dissociation rate ($k_{\text{off}} \sim 0.02 \text{ s}^{-1}$, corresponding to a $t_{1/2}$ of $\sim 24.5 \text{ s}$), and relatively slow association rate ($k_{\text{on}} \sim 3,100 \text{ M}^{-1} \text{ s}^{-1}$), giving a calculated equilibrium constant, K_d , of $6.5 \times 10^{-6} \text{ M}$. But because such measurements are subject to a variety of artefacts and potential misinterpretations⁴, it is important that internal tests of consistency, as well as measurement of parameters in the opposite orientation (MHC/peptide on the solid phase, TCR in solution) lead to essentially the same results.

The authors then evaluated a set of four synthetic variant peptides that have been previously characterized with respect to their agonist or antagonist activity with mature peripheral T cells, as well as in the fetal-thymic-organ culture assay. Three peptides, known to form positively selecting ligands for T-cell precursors bearing the same TCR, form MHC/peptide complexes that bind the TCR less well than those formed with the full agonist. Variant peptides that on binding the MHC fail either to activate mature T cells, or to exert positive selection in the transgenic fetal-thymic-organ culture assay, have k_{off} values that exceed the discriminatory capacity of the method. Surprisingly, the apparent window of affinity that distinguishes the agonist peptide that causes negative selection in the thymic assay from the antagonist peptides that induce both positive selection and act as antagonists for the T-cell clone *in vitro* is quite narrow — the antagonist peptides generate complexes that have roughly two to ten times less affinity for the TCR than does the agonist.

Alam and colleagues' results complement those previously published evaluating a set of agonists in a different TCR-MHC/peptide system⁵. They lead to the neat conclusion, summarized in the figure, that correlates affinity with degree of agonism as well as with the outcome of thymic selection. These data are consistent with kinetic models that view the result of TCR-mediated signals as resulting from differential occupancy^{6,7}; and they tend to count against mechanisms that depend on structural changes, either in conformation or topology⁸, of the TCR that transmit differences in the ligand.

But we should not settle too complacently for this direct affinity/outcome model. In isolation, the new measurements have not integrated the contribution of the CD8 co-receptor, a T-cell surface molecule that not only provides a signature for MHC class I restricted T cells, but also provides



Correlation of activity of MHC/peptide complexes as ligands for peripheral T-cell activation with the outcome of thymic education. Examples of agonist, antagonist or null MHC/peptide ligand displayed on the surface of the antigen-presenting cell (APC) are illustrated. The MHC class I/peptide complex, consisting of a heavy chain (red), a light chain $\beta 2$ -microglobulin (pale blue) and a peptide (dark blue), interacts with the $\alpha\beta$ T-cell receptor (TCR, green and yellow) to initiate positive or negative selection in the thymus, or agonist or antagonist effects on peripheral mature T cells. TCR-MHC/peptide interactions tend to be of moderate affinity. But the highest ones seem to be those in which the agonist peptide participates, which in the case studied by Alam *et al.*¹ are characterized by a dissociation rate constant (k_{off}) of $\sim 0.02 \text{ s}^{-1}$. Antagonist peptides seem to produce complexes of weaker affinity for the TCR (indicated here by greater distance between the APC and T-cell membrane) and null peptides produce MHC/peptide complexes that do not detectably interact with the TCR.

issue¹, Alam *et al.* provide quantitative data on a central issue for immunologists: how the fate of developing T lymphocytes is decided.

The background to the problem is as follows. T lymphocytes undergo a complex 'educational' programme of cellular and molecular development as they pass through the thymus, before being dispatched to peripheral tissues where they sit poised to initiate and control the immune response to infection or to dysregulated protein expression. Both the developmental programme of the T cell and the specificity of its action in recognizing aberrant molecular structures rest on the interaction of clonally distinct T-cell receptors (TCR) with cell-surface molecular complexes composed of major histocompatibility complex (MHC) molecules bound to antigenic or to self peptides on 'antigen-presenting cells' (APC).

peptide was varied. The TCR was cloned from a mature T cell that was activated by an MHC class I molecule (H-2K^b) bound to an octamer peptide derived from chicken ovalbumin. Variants of the ovalbumin peptide have been previously characterized not only for their effects as agonist and antagonist in *in vitro* T-cell activation assays, but also for their effects on positive selection of thymic cells from animals transgenic for the same receptor^{2,3}. Agonist peptides bind H-2K^b and deliver a stimulatory signal to the T cell resulting in sensitization for cytotoxicity; antagonists deliver an inhibitory signal that counters agonist-mediated activation. Positive selection is assessed by exposure of fetal thymic cells in organ culture to synthetic peptides; and by measurement, in an appropriate genetic background, of the development of T cells bearing the transgenic receptor as well as the CD8 marker

additional binding strength for T-cell activation with distinct dynamics⁹. We might expect the effect of CD8 to be to displace the threshold, or even qualitatively to influence the effects of some peptides more than others. The number of peptides sampled for these measurements is small, and consistent outliers, such as one reported for a TCR specific for a self-peptide bound to the MHC class I molecule H-2L^d (ref. 10), may be a warning for those who hold too firmly to such simple models. These affinity measurements are reported for purified molecules, expressed in *Drosophila* or *Pichia* cells (TCR) or in bacteria (H-2K^b), analysed in a binding system with three-dimensional freedom of diffusion: events that take place on cell membranes, constrained to diffusion in two dimensions, may be reflected only indirectly in such free-solution measurements¹¹. Finally, although the TCR analysed in these studies has been tested in a transgenic system, and in fetal-thymic-organ culture, for its susceptibility to peptide-induced positive and negative selection, we should not overlook the artificial nature of the fully selected TCR expressed as a transgene.

Nonetheless, quantitative studies of the

ligand/receptor interactions that regulate immune responses offer objective measurement of parameters that probably reflect molecular controls on the education, selection and function of T lymphocytes. We need only wait for all the right experiments. □

David H. Margulies is in the Molecular Biology Section, Laboratory of Immunology, National Institute of Allergy and Infectious Diseases, National Institutes of Health, Bethesda, Maryland 20892-1892, USA.

1. Alam, S. M. *et al.* *Nature* **381**, 616–620 (1996).
2. Jameson, S. C., Carbone, F. R. & Bevan, M. J. *J. exp. Med.* **177**, 1541–1550 (1993).
3. Hogquist, K. A., Jameson, S. C., Heath, W. R., Howard, J. L. & Bevan, M. J. *Cell* **76**, 17–27 (1994).
4. Schuck, P. *Biophys. J.* **70**, 1230–1249 (1996).
5. Matsui, K., Boniface, J. J., Steffner, P., Reay, P. A. & Davis, M. M. *Proc. natn. Acad. Sci. U.S.A.* **91**, 12862–12866 (1994).
6. McKeithan, T. W. *Proc. natn. Acad. Sci. U.S.A.* **92**, 5042–5046 (1995).
7. Rabinowitz, J. D., Beeson, C., Lyons, D. S., Davis, M. M. & McConnell, H. M. *Proc. natn. Acad. Sci. U.S.A.* **93**, 1401–1405 (1996).
8. Janeway, C. A. *Immun. Today* **16**, 223–225 (1995).
9. Luescher, I. F. *et al.* *Nature* **373**, 353–356 (1995).
10. Al-Ramadi, B. K., Jelonek, M. T., Boyd, L. F., Margulies, D. H. & Bothwell, A. L. M. *J. Immun.* **155**, 662–673 (1995).
11. Dustin, M. L., Ferguson, L. M., Chan, P. Y., Springer, T. A. & Golan, D. E. *J. Cell Biol.* **132**, 465–474 (1996).

SUPERCONDUCTIVITY

Counting the ten-year returns

Paul Grant

HOUSTON, oil capital of Texas, is not exactly New York City, and there was still a year remaining before the first decennial of that incredible rock concert of physics, the 1987 March Meeting of the American Physical Society, christened “another Woodstock” by the late Michael Schluter of AT&T Bell Labs. Why then, asked several journalists in attendance at a meeting in Texas earlier this year*, were we jumping the gun in both time and location? Well, actually the timing was spot on.

It was in the early months of 1986 that Georg Bednorz and K. Alex Müller of the IBM Zürich Research Laboratory caught the first glimpse of their great discovery: a startling drop in resistivity near 30 K in a mixed phase cuprate. As to location, the Houston venue was also eminently fitting, because soon after the appearance of the IBM paper, Paul Chu and his co-workers, both in their home Texas institution and in Alabama, embarked on the series of compositional variations that broke through the liquid nitrogen ‘glass ceiling’ and led to the subsequent sensational events.

The plenary, invited and poster sessions addressed three primary topics in high-temperature superconductivity — pairing mechanisms (both theoretical and experi-

mental aspects), materials and applications. I will concentrate on the third topic, emerging applications. What better tenth anniversary present to Bednorz, Müller and Chu?

The general tone was set in the plenary address by John Rowell, not known from his past presentations for giving an overly sanguine view of the commercial prospects of superconductivity, high- or low- T_c . This time, however, Rowell was quite upbeat, especially regarding high-temperature superconductor (HTSC) electronics, pointing out that a number of price/performance targets set by receiving industries had been met, especially in the case of passive radio frequency filters for cellular or ‘personal communication’ ground stations. The remaining challenge, in his view, was one of perception, not technology, especially regarding the requirement for refrigeration, which is seen by many potential customers as inherently expensive and unreliable. Education and packaging approaches are needed that will make users

as comfortable with the presence of cryogenic devices as the general public is with household refrigerators. Rowell concluded that the next ten years will be “the decade of markets” for superconducting devices, with the prospect of billion-dollar revenues.

In my own speciality, power applications, we announced the successful construction and testing of a 50-metre underground a.c. transmission cable (Aldo Bolza, Pirelli Cavi, Milan). The conductor assembly contained more than six kilometres of lead-stabilized BSCCO tape (BSCCO is a cuprate perovskite structure joined to oxide complexes of bismuth, strontium and calcium). Immersed in liquid nitrogen, the flexible assembly carried a d.c. current of 1,800 A, which is more than twice its design target capacity and established a new record for HTSC current-times-length performance (the record has been broken again since the meeting). In my view, however, what is most important is that the tape was wound on its one-inch diameter hollow core by a standard cable-stranding tool, and not by a team of expensive PhD manual labourers.

An announcement containing a similar pleasant surprise came in the next paper. A collaboration has constructed and operated a 125-horsepower (93 kW) synchronous motor, whose rotor was wound with the same sort of BSCCO tape used in the cable conductor assembly (Dave Driscoll, Reliance Electric). But because the magnetic fields encountered in motor applications are much higher than in cables, the operating temperature had to be 27 K instead of 77 K to achieve the necessary current level. The good news is that the prototype actually produced 200 horsepower, a figure limited by the capacity of the test brake employed and not the intrinsic power of the motor, which is estimated at perhaps 400 horsepower.

Rapid and dramatic developments in BSCCO tape performance have occurred in the past five or six years to make these



A record-breaking power cable. In March, this 50-metre superconducting assembly carried 1,800 A, almost twice as much as an ordinary copper conductor of similar diameter could.

*The 10th Anniversary HTS Workshop on Physics, Materials and Applications, Houston, Texas, USA, 12–16 March 1996.

and other application prototypes possible (Alex Malozemoff, American Superconductor). American Superconductor can produce 10 km of multi-filament BSCCO tape per month, in 1,000-m lengths. Critical current density (J_c) in these tapes can now exceed $44,000 \text{ A cm}^{-2}$. Malozemoff noted that, since 1990, BSCCO wire performance has followed a consistently linear upward trend with time.

Where will it stop? Small regions of BSCCO embedded in the tape, especially near the silver-cladding interface, have been observed magneto-optically to have $J_c > 10^5 \text{ A cm}^{-2}$ at 77 K (David Larbalestier, Univ. Wisconsin). Beyond this limit, and especially to service high-magnetic-field applications at liquid-nitrogen temperatures, a new technology will be needed.

One is already on the horizon (see my earlier News and Views article, *Nature* **375**, 107–108; 1995). Workers from the Los Alamos and Oak Ridge National Laboratories revealed their latest results for biaxially oriented coatings of YBCO (yttrium–barium–copper oxide) on flexible metal tape substrates. The Los Alamos group produces substrate texturization with an auxiliary ion gun during deposition of a buffer film between the metal tape and the YBCO, whereas Oak Ridge directly textures the metal tape using a combination of metallurgical deformation techniques. Texturization is important in producing low-angle grain boundaries and thus good inter-grain coupling. Los Alamos has now observed critical currents of 200 A at liquid-nitrogen temperatures in 2- μm -thick YBCO in short lengths 1 cm wide, a J_c of 10^6 A cm^{-2} .

The Oak Ridge group principally described the textural properties of their method, which at the time of the Houston meeting had produced a J_c of only around 10^5 A cm^{-2} . But while writing this review, I've heard that portions of their samples now have $J_c \approx 5 \times 10^5 \text{ A cm}^{-2}$, an important improvement given that metallurgical texturization may be scaled up to manufacturing processes more easily than ion beam bombardment.

In the meantime, the improving performance of 'first generation' BSCCO wire seems certainly enough to satisfy a number of anticipated applications in the next five years, especially transmission cable. The principal uncertainties are cost and market acceptance of this new technology. The generally acknowledged commercial target for cost/performance is $\$10 \text{ kA}^{-1} \text{ m}^{-1}$, but the latter factor, acceptance, is by far the greater challenge. If the next decade is to see deployment of HTSC wire in power applications, as much attention, if not more, needs to be given to this area as to improving the technology itself.

The workshop was a major event for all of us in the high- T_c trade. As a Hollywood columnist might have put it, it was a bash

where almost everybody who's anybody participated. The next great event will be the American Physical Society March Meeting in 1997 in Kansas City, bringing with it the tenth anniversary of our Woodstock. Unfortunately, this celebration will have to be held somewhere within the confines of a midwest conven-

tion centre, not in the Grand Ballroom of the New York Hilton where it all began. What a pity. □

Paul Grant is at the Electric Power Research Institute, PO Box 10412, 3412 Hillview Avenue, Palo Alto, California 94303, USA.

VISUAL PERCEPTION

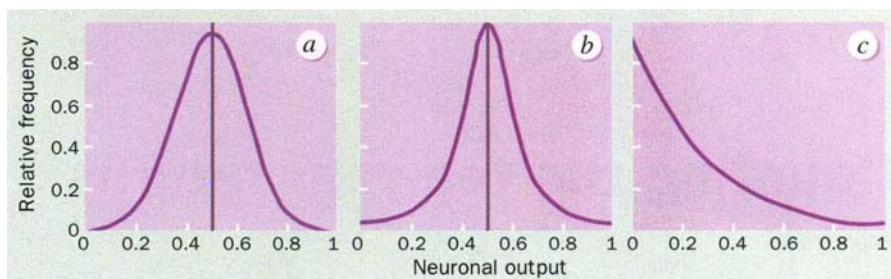
An efficient code in V1?

Roland Baddeley

THE primary visual cortex (V1) is the first cortical area to receive visual input from the eyes. For over 30 years it has been known that the neurons in this part of the brain are sensitive to bars of light of various orientations and sizes^{1,2}, and this phenomenon constitutes a common, crucial stage of visual perception in mammals. What has been far less clear is why neurons in V1 show such selectivity.

neighbouring measurements. Therefore, by using such a code, efficient transmission is achieved because the intensity information is preserved while allowing the minimization of some measure of neuronal output capacity such as total variance.

Several researchers have explored this notion of efficiency. By using estimates of the correlation in image intensity in natural scenes as a function of distance, character-



What is a sparse representation? For a system with binary outputs, sparsity can be simply equated with a low average firing probability. For a system with continuous outputs, a definition is more difficult. But one method is to define a measure of how 'peaked' the distribution of outputs is around its mean. Consider the distribution of outputs from a single 'filter' when presented with a large collection of natural images. If the filter has random spatial weightings, then by the central limit theorem we would expect the distribution of outputs to be gaussian (a). If, on the other hand, the weightings of the filter matched some nonrandom structure in the images, we may find a non-gaussian output distribution. One simple way a filter's output distribution could differ from gaussian is to be more concentrated around its mean value (b), and a representation with an output distribution that resembles b rather than a is called 'sparse' (see the paper under discussion³). Another reason for having such filters is that they minimize the output entropy for a fixed output variance³. Given the more biologically plausible constraint on the average firing rate of the neuron (and hence metabolic cost), one particular sparse output distribution (c) maximizes output entropy.

On page 607 of this issue³, Olshausen and Field describe some simulations that shed light on this question. They show that, for an appropriate definition of 'efficiency', the response properties of neurons in V1 qualitatively (and in some aspects quantitatively) resemble the most efficient linear representation of natural images possible.

Their definition of efficiency is probably best understood by reference to the coding performed by a more peripheral part of the visual system — the retinal ganglion cells that provide the output from the eye to the brain. Rather than simply signalling raw local image intensity, these cells appear to signal the difference between the intensity at a given location and that predicted from neighbouring locations. Image intensity in natural scenes is highly predictable from

izing potential sources of noise, and formalizing efficiency (for example using information theory), predictions of the precise form of processing can be made. Such an approach has led to accurate quantitative predictions of contrast sensitivity^{4,5}, colour sensitivity⁶ and the form of inhibition in fly retina⁷.

Unfortunately, early attempts to apply the same notion to the coding performed by cells in the primary visual cortex have been less successful. One method that has had some popularity is a form of factor analysis, known as principal components analysis (PCA), which finds the linear combinations of image intensities (components) that capture as much of the variation in natural images using as few components as possible. Although there

are some similarities between the neuronal properties predicted by this method and those observed empirically (the lower order components are well described as spatial and orientation derivatives of various orders⁸), it is far from adequate as a general model. The resulting spatial weightings only vaguely resemble those observed in V1 and, more importantly, PCA is only appropriate if the number of components is far less than the number of inputs (the number of neurons in V1 is much larger than the number of cells from the lateral geniculate nucleus providing input). Accounting for the variation in images alone is an insufficient constraint on coding to generate the properties of cells in V1.

What Olshausen and Field now show is that if, in addition to accounting for the variations in the image, representations are also constrained to be 'sparse' (a given image is represented by only a few 'active' cells out of a potentially much larger number — see figure), then the optimal representation of natural scenes possesses many similarities to the representation found in V1. The simulations reported provide an impressive existence proof that simple learning rules can be used to determine receptive-field-like structures. With an intuitive definition of efficiency, optimized using a simple neural-network learning rule, the representations generated share many qualitative similarities to those found in V1.

The question then becomes why these principles should underlie the processing performed. Olshausen and Field argue that there are two reasons: sparsity could be a signal that 'interesting' image structure has been identified; and sparse representations may be better suited for further processing, independent of whether or not they represent underlying input regularities, because they minimize output entropy for a given output variance.

Two comments seem appropriate. First, although use of characteristics of the output distribution (for example, sparsity) is an accepted statistical method for identifying structure⁹, and Olshausen and Field present three simulations showing that optimizing sparsity can indeed identify the desired structure, sparsity alone

does not necessarily identify 'interesting' structure because very sparse representations can be found for trivial reasons. Specifically, if the local intensity variance is not constant, then any filter with spatial weightings that sum to zero will generate a sparse representation¹⁰.

Second, it is not clear that output variance is an appropriate measure of neuronal channel capacity, as Olshausen and Field implicitly argue when they propose that sparse codes minimize output entropy. Rather, because the brain uses a large amount of energy, and V1 is the largest visual area in the brain, any strategy that would minimize energy use would be useful. Instead of constraining neurons by output variance, it seems more appropriate to

constrain their average firing rate. A code that maximizes output entropy when constrained by a given firing rate will be sparse — in fact the distribution of output states will be exponential¹¹. One of the proposed measures of 'sparsity' measured average firing rate ($|x|$), and this, it is claimed, will also generate V1-like receptive fields. So a potentially simpler explanation of Olshausen and Field's results is that the representations of the world found in V1 are 'efficient' simply because such representations will consume less energy. □

Roland Baddeley is in the Departments of Experimental Psychology and Physiology, University of Oxford, South Parks Road, Oxford OX1 3UD, UK.

GROWTH FACTORS

Mad connection to the nucleus

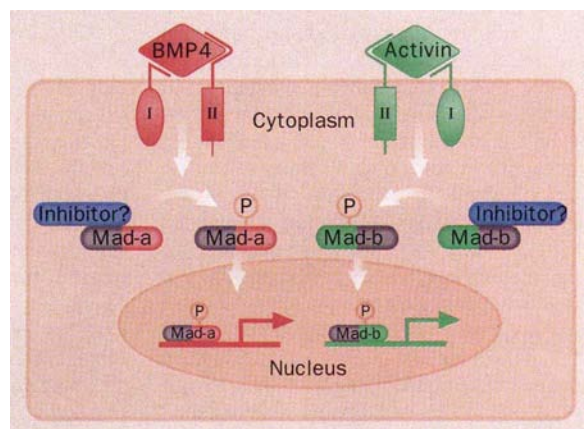
Christof Niehrs

MEMBERS of the transforming growth factor- β (TGF- β) family of proteins have a large part in cell signalling in animals. Expansion of genes encoding them was probably a milestone in vertebrate evolution, for they carry out such a wide variety of duties in development: they regulate the responses of some cells to other growth factors; they regulate cell growth and differentiation; and they act as inductive signals in the embryo. Very little is known about TGF- β signal transduction from the cell membrane to the nucleus, but three new papers present insights into the process. On page 620 of this issue¹, Liu *et al.* identify so-called *Mad* genes as encoding a previously unknown class of transcription factors which act downstream of TGF- β signals. And in *Cell*^{2,3}, two groups describe fruitful investigations into the regulation and specificity of *Mad* protein function.

About 40 members of the TGF- β gene superfamily are known to encode secreted proteins. These form homo- or heterodimers, so total TGF- β ligand diversity is high. The receptors for TGF- β ligands consist of complexes of a type I and a type II transmembrane serine/threonine kinase. There are a number of type I and type II isoforms that could by combination account for considerable receptor diversity⁴. Probably as a result of this

ligand/receptor variety, a single cell seems to respond differently when challenged with different concentrations or types of TGF- β proteins.

For example, the product of the TGF- β -type *Drosophila* gene *decapentaplegic* (*dpp*) acts as a morphogen in patterning of the embryonic wing, where it can elicit



Model of the *Mad* signal-transduction pathway in TGF- β signalling. A TGF- β ligand, here BMP4 and activin, binds and activates the transmembrane serine/threonine kinase of its corresponding type I/II receptor complex⁴. Either directly or by an unknown kinase cascade, this event activates the phosphorylation of a ligand-matched *Mad* protein in the cytoplasm. Phosphorylation enables *Mad* to move into the nucleus, which may also require dissociation of a hypothetical inhibitory protein interacting with the amino-terminal *Mad* domain (black) or direct unmasking of a nuclear localization signal. Once in the nucleus, *Mad* mediates transcriptional activation of specific genes.

various responses at various doses. Thus, an individual cell appears to distinguish different concentrations of *dpp* protein and respond accordingly (as discussed in News and Views last month⁵). Similar observations have been made in *Xenopus* embryos, where the animal cap system is

- Hubel, D. H. & Wiesel, T. N. *J. Physiol., Lond.* **148**, 574–591 (1959).
- Hubel, D. News and Views *Nature* **380**, 197–198 (1996).
- Olshausen, B. A. & Field, D. J. *Nature* **381**, 607–609 (1996).
- Atick, J. J. & Redlich, A. N. *Neural Comp.* **4**, 196–210 (1992).
- Van Hateren, J. H. *Vision Res.* **33**, 257–267 (1993).
- Atick, J. J., Li, Z. & Redlich, A. N. *Neural Comp.* **4**, 559–572 (1992).
- Srinivasan, M. V., Laughlin, S. B. & Dubs, A. *Proc. R. Soc. B* **216**, 427–459 (1982).
- Baddeley, R. J. & Hancock, P. J. B. *Proc. R. Soc. B* **246**, 219–223 (1991).
- Huber, P. *J. Ann. Statist.* **13**, 435–475 (1985).
- Baddeley, R. J. *Network* **7**, 409–421 (1996).
- Shannon, C. E. & Weaver, W. *The Mathematical Theory of Communication* (Univ. Illinois Press, Urbana, IL, 1949).

used to study inductive processes. Animal cap cells can mount a bewildering range of responses to TGF- β proteins. For instance, a member of the bone-morphogenetic-protein subfamily (BMP4) can induce ventral-type mesoderm only, *Xnr3* induces neural tissue, and different doses of activin induce at least four different mesodermal cell types. How, then, is signalling by different TGF- β ligands and doses translated into distinct transcriptional responses?

Although different receptor isoforms may confer specificity at the membrane level, it has been unclear how specificity is maintained during signal transduction to the nucleus. In *Xenopus*, mesoderm induction by activin, fibroblast growth factor and BMP4 all depend on Ras function, which doesn't help to further our understanding of specificity.

An important clue towards identifying the downstream effectors of TGF- β -type signals came from a genetic screen of *Drosophila* by Gelbart and colleagues⁶, who systematically searched for enhancers of a *dpp* mutant phenotype. They identified *Mothers against dpp* (*Mad*), a gene that turned out to be a member of a family of related genes in *Caenorhabditis elegans*, *Sma2-4* (ref. 7). *Mad* family genes encode proteins of relative molecular mass about 50K. They contain conserved amino- and carboxy-terminal domains separated by a proline-rich linker, but they do not have any protein motif that gives a clue to their function.

With the aid of a database of expressed sequence tags, Liu *et al.*¹ have now cloned *Smad1*, a human homologue of *Mad*. As described in the paper, they used the GAL4-system to find out if the gene encodes a transcription factor: it emerged that the carboxy-terminal domain of *Smad1* indeed activates transcription, but the full-length protein does not. When cells are cotransfected with BMP type I and II receptors, however, and then treated with BMP4 ligand, full-length *Smad1* becomes transcriptionally active. This implies that *Mad* proteins have latent transcription-factor activity that is activated by ligand binding to TGF- β receptors. So *Mad* proteins may provide the link in signal transduction from the cytoplasm to the nucleus in response to TGF- β growth factors.

It remains to be seen whether *Mad* proteins can bind DNA, but from the new results it seems that they can directly initiate transcriptional responses. Because *Smad1* protein enters the cell nucleus in response to BMP4, *Mad* latency may be due to its retention in the cytoplasm. Regulation of transcription factors by controlling their sojourn in the cytoplasm is a recurring strategy in signal transduction⁸, and an attractive hypothesis for *Mad* control. The main message of one of the *Cell* papers, that by Hoodless *et al.*², is that phosphory-

lation of the protein may be the event that instructs it to enter the nucleus.

This leaves the issue of explaining the specificity of signal transduction by different TGF- β -type growth factors. Enter Graff *et al.*³, who cloned two *Mad* homologues in *Xenopus*, *Xmad1* and *Xmad2*, and used the animal cap system to show that their protein products mediate distinct TGF- β signalling responses. Like BMP4, *Xmad1* induces ventral mesoderm only; and like medium and high doses of activin, *Xmad2* induces dorsolateral and dorsal mesoderm in a concentration-dependent manner. Signal transduction by a specific set of *Mad* proteins may therefore explain the differential response of individual cells to various TGF- β -type growth factors (see figure). Given that *Mad* proteins elicit such specific responses, they could become a promising pharmaceutical target — not least because the first *Mad* candidate for a tumour-suppressor gene has just surfaced⁹.

We are of course left with many questions. How can overexpression of *Mad* proteins lead to a response when their activity should be latent? Like the transcription factor NF- κ B, whose nuclear translocation is inhibited by I- κ B, retention of *Mad* protein in the cytoplasm may involve accessory proteins that may be titrated (see figure), explaining the relatively high doses of exogenous *Smad1* and *Xmad1* required for induction of ventral mesoderm. If the dose-dependent effects of TGF- β really are transduced by individual, 'smart' *Mad* proteins, how do they integrate the signal dose? Finally, what are the target genes for *Mad* and what determines their specificity? We can expect a few of the answers before too long — indeed, some of them are undoubtedly on still-wet autoradiograms in the many laboratories working in this area. □

Christof Niehrs is in the Division of Molecular Embryology, Deutsches Krebsforschungszentrum, INF 280, 69120 Heidelberg, Germany.

1. Liu, F. *et al.* *Nature* **381**, 620–623 (1996).
2. Hoodless, P. A. *et al.* *Cell* **85**, 489–500 (1996).
3. Graff, J. M., Bansal, A. & Melton, D. A. *Cell* **85**, 479–487 (1996).
4. Kingsley, D. M. *Genes Dev.* **8**, 133–146 (1994).
5. Smith, J. *Nature* **381**, 367–368 (1996).
6. Sekelsky, J. J., Newfeld, S. J., Raftery, L. A., Chartoff, E. H. & Gelbart, W. M. *Genetics* **139**, 1347–1358 (1995).
7. Savage, C. *et al.* *Proc. natn. Acad. Sci. U.S.A.* **93**, 790–794 (1996).
8. Vandromme, M., Gauthier-Rouviere, C., Lamb, N. & Fernandez, A. *Trends biochem. Sci.* **21**, 59–64 (1996).
9. Hahn, S. A. *et al.* *Science* **271**, 350–353 (1996).

Erratum

In the article 'Continents of the core' by Michael E. Wyssession, published on 30 May (*Nature* **381**, 373–375; 1996), the last reference was inadvertently left out. It should have read:

7. Kingma, K. J., Cohen, R. E., Hemley, R. J. & Mao, H.-K. *Nature* **374**, 243–245 (1995).

DAEDALUS

Fire from heaven

BENJAMIN FRANKLIN famously proved that lightning was electrical, by flying a kite in a thunderstorm and drawing sparks from the wet string onto his knuckle. He was lucky. The next man to try it was electrocuted. Indeed, the test could easily provoke a lightning strike. The deadliest version of the experiment, says Daedalus, would be to fly the kite on a string impregnated with caesium or its salts. The resulting discharge would follow the string to Earth, boiling off its caesium, whose vaporized ions conduct electricity very readily. The resulting column of hot conducting plasma might act as a sort of extended lightning. All the energy of the thunderstorm would drain to Earth in one long, roaring blaze.

This could be a neat way of short-circuiting a thunderstorm. The farmer seeing a downpour about to wreck his harvest, the sportsmen facing the washing out of their match, the anxious outdoor concert promoter, all might love to short out a gathering storm and get it over with. But what heroic fool would fly the deadly initiating kite? Daedalus has a safer idea. A rocket powered by caesium-rich fuel could be fired into the clouds by remote control. It would leave a hot, conducting trail, similar to the ionized exhaust plume by which he planned last week to power an electrically augmented space rocket. With one deafening blast, the thunderstorm would drain to Earth down the trail of the rocket. The clouds, deprived of their mechanism for electrostatic precipitation, would drift away harmlessly.

But a thunderstorm is not an isolated event. The world's thunderstorms keep the lower ionosphere, 40 km up, charged to over 200 kV. So Daedalus is scaling up his ideas. He is planning a big rocket, powered by a mixture of caesium perchlorate and buckminsterfullerene (both of which form ions very readily), to soar into the ionosphere and short it out.

This stable 'cable' will draw power from the whole conducting ionospheric shell. Its fiery trail to Earth will be a sort of permanent lightning, compacted and stabilized by the magnetic pinch effect of its steady current. With a caesium-rich ground electrode to accept and anchor the discharge, and release ions to keep it going, it will deliver gigawatts of power to the distribution grid. Sadly, the world's thunderstorms will not be arrested, but intensified — to return to the ionosphere the huge current being drawn from it. The column could also usefully serve as a global transmitter. Radio and television signals sent up it would be conducted round the Earth on the ionosphere, and radiated down to receivers everywhere.

David Jones

Preclinical test for prion diseases

SIR — The recent recognition that there may be a link between bovine spongiform encephalopathy (BSE) in cattle and Creutzfeldt-Jakob disease (CJD) in humans¹ highlights the urgent need to find a method for the preclinical diagnosis of transmissible spongiform encephalopathies, or prion diseases. Until then, given the long incubation period of the disease agent, control measures will be hampered. Here we suggest a possible approach to an early preclinical diagnostic method for transmissible spongiform encephalopathies using scrapie in sheep as a model.

Infected animals and humans have neither a disease-specific immune response nor consistent biochemical, haematological and gross pathological abnormalities. The early diagnosis of transmissible spongiform encephalopathies therefore depends on the recognition of clinical signs, electroencephalography or magnetic resonance imaging techniques, or the more invasive method of taking brain biopsies.

Confirmatory diagnosis, however, depends on histological examination of the brain of suspected cases. An altered protein, PrP^{Sc}, can be detected in the brain of diseased individuals, by immunohistochemistry or other protein-detection

methods. Numerous studies have confirmed PrP^{Sc} as a sensitive and specific marker of this group of diseases².

Immunohistochemistry does not demonstrate infectivity, but it does have many advantages over the mouse bioassay, which is considered a more sensitive detection method but far too cumbersome and time-consuming to become a practical diagnostic method. Various tissues — blood, urine, tissue fibroblasts and, particularly in animals, lymphoid tissue — have been used in attempts to develop an early-detection technique for these encephalopathies. Most attempts were either negative or inconclusive³, but we have built on the classic work of Hadlow, who showed that in sheep naturally infected with scrapie, infectivity was detectable in lymphoid tissue as early as 10–14 months of age, well before it occurred in the central nervous system^{4,5}.

Having demonstrated the consistent presence of the scrapie-associated PrP^{Sc} in tonsils of a group ($n=55$) of naturally infected, clinically positive scrapie sheep by immunohistochemistry⁶, using antibodies directed to selected, synthetic PrP-based peptides⁷, we have now detected PrP^{Sc} in tonsils of sheep in the preclinical stage of the disease, long before the onset of clinical signs.

We selected a group of ten purposely bred lambs, six of them homozygous for the PrP^{VQ} allele (with the residue valine (V) at position 136 of the amino-acid sequence and glutamine (Q) at position 171). The remaining four lambs were heterozygous, possessing one PrP^{VQ} allele and one PrP^{AR} allele (with alanine at position 136 and arginine at position 171). In several breeds, this PrP^{VQ} allele is significantly associated with an increased susceptibility to scrapie, whereas the PrP^{AR} allele is significantly associated with increased resistance of sheep to scrapie⁸.

The ten lambs were born and raised on a farm where scrapie has been occurring for several years. In this flock, we had observed that sheep with the genotype PrP^{VQ/VQ} died from scrapie when about 25 months old, but that most of the sheep with the genotype PrP^{VQ/AR} were still healthy at 70 months. Born

and raised in this environment, the PrP^{VQ/VQ} sheep would thus be expected to develop disease symptoms when about 25 months old, whereas the PrP^{VQ/AR} sheep would stay healthy. We therefore regarded these two groups of sheep as a suitable model for studying changes at known stages of the incubation period.

We took tonsillar biopsies from these animals at approximately 10 (range 9.5–10) months after birth, when none of the sheep showed clinical signs of scrapie. We found clear, extensive PrP^{Sc} immunostaining in all six susceptible PrP^{VQ/VQ} sheep (*a* in the figure) at 10 months of age, but detected no PrP^{Sc} immunostaining in any of the resistant PrP^{VQ/AR} sheep (*b* in the figure).

We have thus detected scrapie-associated PrP^{Sc} in tonsils of sheep at less than half the expected incubation period, approximately one year before the expected onset of clinical disease. We did not detect this PrP^{Sc} protein in sheep expected to develop scrapie at a much later stage or to remain healthy throughout their lives. Thus, screening tonsillar tissue for PrP^{Sc} by immunohistochemistry offers a possibility of preclinical diagnosis in sheep scrapie.

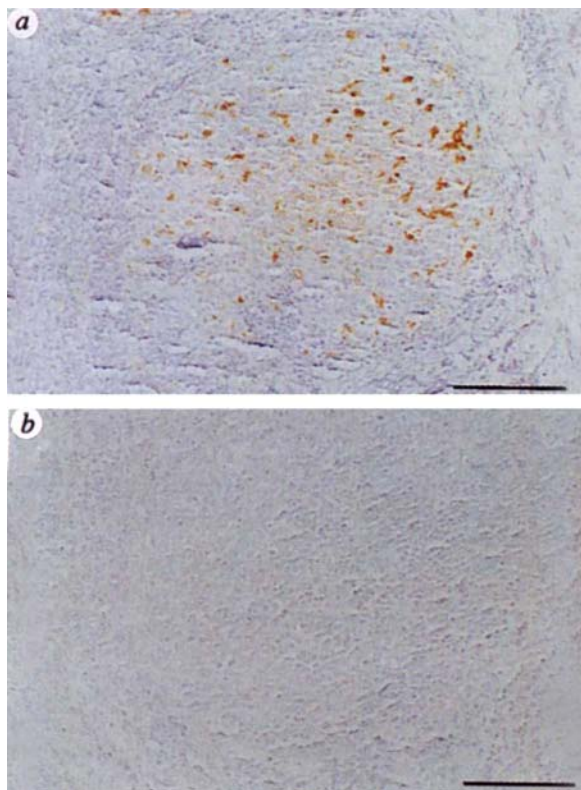
In view of the negative transmission experiments using different peripheral tissues of cattle affected by BSE⁹, our technique may not work in cattle. However, because immunohistochemistry has not been tried in depth as a diagnostic test, and because the mouse bioassay may not be sensitive enough in the case of BSE, it is surely worth investigating our suggestion further.

An early diagnosis of BSE could alleviate the draconian measures proposed in the United Kingdom for culling potentially BSE-infected cattle. The technique of taking tonsillar biopsies in live cattle is feasible and would be even easier than in sheep. In human spongiform encephalopathies, this approach has even brighter prospects, as infectivity has been detected in various lymphoid tissues of CJD patients¹⁰, and tonsil biopsy is a relatively noninvasive, easy technique.

**B. E. C. Schreuder, L. J. M. van Keulen
M. E. W. Vromans, J. P. M. Langeveld
M. A. Smits**

*DLO-Institute for Animal Science
and Health (ID-DLO), PO Box 65,
8200 AB Lelystad, The Netherlands*

- Will, R. G. *et al.* *Lancet* **374**, 921–925 (1996).
- Prusiner, S., Collinge, J. & Anderton, B. in *Prion Diseases in Humans and Animals* (Ellis Horwood, London, 1992).
- Schreuder, B. E. C. *Vet. Q.* **16**, 174–181 (1994).
- Hadlow, W. J., Kennedy, R. C. & Race, R. E. *J. infect. Dis.* **146**, 657–664 (1982).
- Hadlow, W. J. *et al.* *Vet. Path.* **17**, 187–199 (1980).
- van Keulen, L. J. M. *et al.* *J. clin. Microbiol.* **34**, 1228–1231 (1996).
- van Keulen, L. J. M. *et al.* *Vet. Path.* **32**, 299–308 (1995).
- Belt, P. B. G. M. *et al.* *J. gen. Virol.* **76**, 509–517 (1995).
- Fraser, H. & Foster, J. D. in *Proc. Consult. BSE sci. vet. Committee CEC 145–160* (Brussels, September 1993).
- Brown, P. *et al.* *Ann. Neurol.* **35**, 513–519 (1994).



a, Positive PrP^{Sc} immuno-staining in lymphoid follicles in tonsillar biopsy from a scrapie-susceptible sheep (PrP^{VQ/VQ}) at age 10 months. Peroxidase-labelled streptavidin-biotin staining. *b*, Negatively staining tonsillar biopsy from a PrP^{VQ/AR} sheep at the same age. Scale bars, 100 μ m.

Forest canopy productivity index

SIR — Increased carbon dioxide concentrations can stimulate the productivity of terrestrial vegetation and thereby alter the relationships between anthropogenic carbon emissions, atmospheric CO₂ concentrations and climate change. This must therefore be taken into account when policy options for climate control are being assessed^{1,2}.

The representation of increased CO₂ on a global scale should come from models, but the growth and carbon storage of different vegetation types in controlled, CO₂-enriched atmospheres must be determined experimentally in order to test assumptions made in such models. Critical mechanistic model components (for example, photosynthesis and transpiration) can be compared with experimental data³, but we do not have enough experimental data on the growth of trees to evaluate models of forest responses to rising CO₂. Results of the few field experiments that have investigated tree responses beyond the seedling stage are difficult to generalize and extrapolate. I suggest that some of this difficulty derives from the use of plant dry mass as the primary indicator of response. A more robust and conceptually more useful measure of response is the annual production of wood per unit of leaf area, which I refer to here as canopy productivity index (CPI; formerly called growth efficiency⁴).

The effect of CO₂ enrichment on plant mass has varied widely in the several studies with broadleaf trees (see table). In 1992, we reported that *Liriodendron tulipifera* trees grown for almost three years in elevated concentrations of atmospheric CO₂ showed a sustained increase in photosynthesis rate compared with trees grown in ambient CO₂, but with no significant increase in plant dry mass⁵, in marked contrast to the reported responses of *Citrus aurantium* trees exposed to elevated CO₂ under horticultural conditions⁶. The difference in response in these two studies has been taken as evidence that trees will show growth responses only in the presence of adequate nutrients⁷, or attributed to

RESPONSE TO ELEVATED CO₂ OF DRY MASS AND ANNUAL STEM PRODUCTION PER UNIT LEAF AREA (CPI) OF FIELD-GROWN BROADLEAF TREES

Species	Duration of exposure (growing seasons)	% Increase in dry mass	% Increase in CPI	Ref.
<i>L. tulipifera</i>	2.7	18	35	4, 5
<i>Populus x euramericana</i> (Eugenei)				7, 11
Low fertility	1	26	22	
High fertility	1	38	18	
<i>Populus x euramericana</i> (Robusta)	2	44	37	12, 13
<i>Populus trichocarpa x deltoides</i> (Beaupré)	2	73	22	12, 13
<i>Fagus sylvatica</i>	2	91	31	14
<i>Q. alba</i>	4	135	37	4, 9
<i>C. aurantium</i>	5	180	33	6, 15
Average ± s.d.		76 ± 57	29 ± 7	

In each experiment the trees were planted directly in the ground and exposed in open-top chambers to CO₂ partial pressures of approximately 350 p.p.m. (ambient) and 650–700 p.p.m. CPIs of *L. tulipifera* and *Q. alba* were calculated by regression analysis of annual stem mass increment versus leaf area. Other calculations were based on published values of mean stem dry mass or dry mass increment (or a surrogate measure), and leaf area or relative increase in leaf area.

unusual leaf physiology in *Citrus*⁸. Such explanations are unnecessary, and the nutrient explanation is not robust. We recently reported that *Quercus alba* saplings exposed to elevated CO₂ for four years responded similarly to the *Citrus* trees, more than doubling in dry mass relative to trees in ambient CO₂, even though they were grown in the same exposure chambers and unfertilized soil as the *L. tulipifera*⁹. Other species have shown intermediate responses (see table).

Despite this apparent divergence of results, which makes generalizations untenable, closer analysis suggests a fundamental similarity in response across these studies. Regardless of the substantial differences in cultural conditions, inherent species characteristics and the final dry mass response, the CPI of broadleaf tree species increases approximately 29% in response to CO₂ concentration increases of 300–350 parts per million (p.p.m.; see table). The large increases in dry mass observed in *Q. alba* and *C. aurantium* were dependent on an increasing leaf area, a response that was established very early in the exposure regime and was compounded over time. Conversely, the limited response of the *L. tulipifera* was attributed to a relative decline in leaf area. The CPI normalizes the increases in wood accumulation to a

constant leaf area, and therefore is more relevant to the response of trees in a forest where leaf area approaches a relatively constant value, constrained by available resources, particularly light, water and nutrients. The 29% increase in CPI is similar to the growth increases observed in many short-term studies with small seedlings, as well as to the amount of CO₂ fertilization required to balance most global carbon budget models^{8,10}.

The consistent response of the CPI to elevated CO₂ suggests that sustained, moderate increases in tree growth can be expected in a CO₂-enriched atmosphere. That is not to say, however, that this index alone predicts forest productivity. Root growth and heterotrophic respiration are not taken into account in the CPI. Environmental factors that reduce the leaf area of a forest stand, such as soil infertility or drought, will reduce productivity gains from the 29% enhancement suggested by the CPI.

This analysis gives rise to several hypotheses about tree responses which can be tested in larger-scale CO₂-enrichment experiments in forests: for example, CPI will remain higher in elevated CO₂ throughout stand development and canopy closure; and effects of other environmental factors on tree growth will be manifested primarily through changes in leaf area index. The separation of functional variables such as leaf physiology and wood production, and structural variables such as leaf area index, is the basic approach of a global productivity model³. The CPI provides a useful construct for organizing existing experimental data and future, large-scale forest exposure experiments along similar principles.

R. J. Norby

Environmental Sciences Division,
Oak Ridge National Laboratory,
Oak Ridge, Tennessee 37831-6422, USA

- Wigley, T. M. L., Richels, R. & Edmonds, J. A. *Nature* **379**, 240–243 (1996).
- Friedlingstein, P. et al. *Global Biogeochem. Cycles* **9**, 541–556 (1995).
- Woodward, F. I., Smith, T. M. & Emanuel, W. R. *Global Biogeochem. Cycles* **9**, 471–490 (1995).
- Norby, R. J., Wullschlegel, S. D. & Gunderson, C. A. in *Carbon Dioxide and Terrestrial Ecosystems* (eds Koch, G. W. & Mooney, H. A.) 1–21 (Academic, San Diego, 1996).
- Norby, R. J., Gunderson, C. A., Wullschlegel, S. D., O'Neill, E. G. & McCracken, M. K. *Nature* **357**, 322–324 (1992).
- Idso, S. B. & Kimball, B. A. *Global Biogeochem. Cycles* **7**, 537–555 (1993).
- Curtis, P. S., Vogel, C. S., Pregitzer, K. S., Zak, D. R. & Teeri, J. A. *New Phytol.* **129**, 253–263 (1995).

- Amthor, J. S. *Global Change Biol.* **1**, 243–274 (1995).
- Norby, R. J., Wullschlegel, S. D., Gunderson, C. A. & Nietch, C. T. *New Phytol.* **131**, 91–97 (1995).
- Wullschlegel, S. D., Post, W. M. & King, A. W. in *Biotic Feedbacks in the Global Climatic System. Will the Warming Feed the Warming?* (eds Woodwell, G. M. & Mackenzie, F. T.) 85–107 (Oxford Univ. Press, New York, 1995).
- Pregitzer, K. S. et al. *New Phytol.* **129**, 579–585 (1995).
- Ceulemans, R., Jiang, X. N. & Shao, B. Y. *Ann. Bot.* **75**, 609–617 (1995).
- Ceulemans, R. et al. *Tree Physiol.* **16**, 61–68 (1996).
- Mousseau, M. et al. in *Carbon Dioxide and Terrestrial Ecosystems* (eds Koch, G. W. & Mooney, H. A.) 71–86 (Academic, San Diego, 1996).
- Idso, S. B., Wall, G. W. & Kimball, B. A. *Envir. exp. Bot.* **33**, 367–375 (1993).

Clutch size and malaria resistance

SIR — It was suggested more than a decade ago that the cost of egg production may be a major factor in the evolution of clutch size in birds¹, and that although this cost is a fundamental assumption of theories on the evolution of life-history patterns, it had not yet been adequately tested^{1,2}. Manipulation of the rate of egg-laying is a prerequisite for demonstrating this cost, but has been considered technically difficult¹. Here we provide the first experimental evidence for a trade-off between egg production and parasite defence in the great tit (*Parus major*), a classic model species in the study of life-history evolution.

We removed the first two eggs laid in half the nests of our breeding population. Females in this experimental group therefore compensated by laying, on average, one more egg than females in a control group (*a* in the figure), but raised, on average, one offspring less (*b* in the figure). Fourteen days after their chicks had hatched, we assessed the prevalence of *Plasmodium* in the peripheral blood-stream of females.

Plasmodium is a malarial parasite that is common in the population of birds we are studying, and which potentially impairs survival and future reproduction^{3,4} of its host. If there is a trade-off between egg production and parasite defence, one

would predict a higher prevalence of parasites in the females of the experimental group, as they laid a larger clutch. Prevalence in males of the experimental group should not increase, because the number of offspring to be raised is no greater than in the control group. We found that an increase in the clutch size by one egg leads to an increase in the prevalence of *Plasmodium* from 20% in the control females to 50% in the experimental females (*c* in the figure), thus supporting the hypothesis of a trade-off between egg production and parasite defence. As expected, there was no difference in the prevalence of *Plasmodium* in males of the two groups (*c* in the figure).

Evolutionary theory predicts that the optimal clutch size is the result of a trade-off between current and future reproduction^{2,5,6}. At the physiological level, this trade-off predicts that parents investing heavily in their current offspring will have fewer resources to allocate to parasite defence, thereby impairing their future reproduction⁷. For birds, the studies in which clutch or brood size were experimentally manipulated after egg-laying have shown that the clutch size giving rise to the most recruits is larger than the most common clutch, but has produced only equivocal evidence for a general trade-off



Heinz Richner

The cost of egg production in birds such as the great tit (*Parus major*) shown here may be an important factor in the evolution of clutch size.

in life-history between current and future reproduction⁸⁻¹¹.

It was stated recently that "all brood size manipulations ignore costs incurred earlier in the life history — in particular the cost of producing eggs, which might well affect the results of brood manipulation experiments" (p.158 of ref. 2). Our results support this contention, and provide evidence that the cost of egg production may be an important factor in the evolution of optimal clutch size in birds.

Anne Oppliger*, Philippe Christe

Zoology Department, IZEA,
University of Lausanne,
CH-1015 Lausanne, Switzerland

Heinz Richner

Zoology Department,
University of Bern,
CH-3032 Hinterkappelen, Switzerland

1. Partridge, L. & Harvey, P. H. *Nature* **316**, 20 (1985).
2. Stearns, S. *The Evolution of Life Histories* (Oxford Univ. Press, 1992).
3. Hayworth, A. M., van Riper, C. III & Weathers, W. W. *J. Parasit.* **73**, 850–853 (1987).
4. Richner, H., Christe, P. & Oppliger, A. *Proc. natn. Acad. Sci. U.S.A.* **92**, 1192–1194 (1995).
5. Williams, G. C. *Am. Nat.* **100**, 687–690 (1966).
6. Charnov, E. L. & Krebs, J. R. *Ibis* **116**, 217–219 (1974).
7. Gustafsson, L. et al. *Phil. Trans. R. Soc. B* **346**, 323–331 (1994).
8. Gustafsson, L. & Sutherland, W. J. *Nature* **335**, 813–815 (1988).
9. Lindén, M. & Möller, A. P. *Trends Ecol. Evol.* **4**, 367–371 (1989).
10. Lessels, C. M. in *Behavioural Ecology* (eds Krebs, J. R. & Davies, N. B.) 32–68 (Blackwell, Oxford, 1991).
11. Pettifor, R. J. *Anim. Ecol.* **62**, 145–159 (1993).
12. Krebs, C. J. *Ecological Methodology* (Harper Collins, New York, 1989).

*The order of the authors was determined by tossing a coin. Correspondence should be addressed to H.R.

Erratum

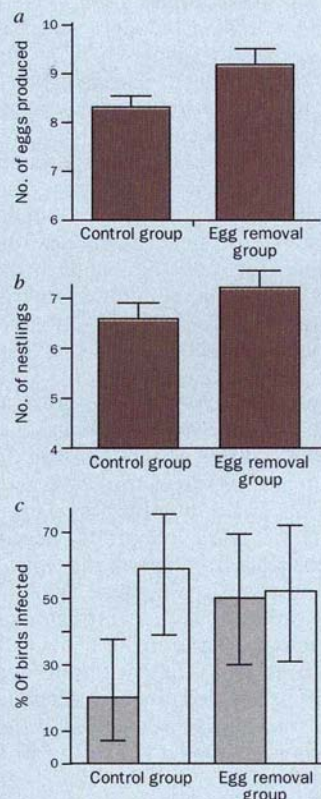
IN the Scientific Correspondence "Roots in mixotrophic algae" published in the 30 May issue (*Nature* **381**, 382; 1996), the authors' affiliations were transposed. They are shown correctly below.

J. R. M. Chisholm*, C. Duga†
E. Ageron†, P. A. D. Grimont†
J. M. Jaubert*

*Observatoire Océanologique Européen,
Centre Scientifique de Monaco,
MC 98000 Principality of Monaco

†Unité des Entérobactéries,
Unité 386 INSERM, Institut Pasteur,
75724 Paris Cedex 15, France

a, Mean number of eggs ($\pm$ s.e.) laid by females of a control group ($n = 41$ broods) and an experimental group in which the first two eggs were removed the day of laying ($n = 35$ broods). Females of the test group produced a larger clutch than females of the control group ($t = 1.99$; 74 d.f.; $P_{2-tailed} = 0.05$). There was no difference between the two groups in mean hatching date (mean difference = 0.5 days; $t = 0.30$; 67 d.f.; $P_{2-tailed} = 0.76$). *b*, Mean number of nestlings ($\pm$ s.e.) 14 days post-hatching in the control group ($n = 37$) and in the test group ($n = 33$). Females of the test group raised significantly fewer chicks than did females of the control group ($t = 3.06$; 68 d.f.; $P_{2-tailed} = 0.003$). *c*, Prevalence of *Plasmodium* spp. ($\pm$ 95% confidence limits of binomial distribution¹²) in male (open bars) and female (shaded bars) great tits in the test group ($n = 30$) and the control group ($n = 35$). Manipulation significantly affected the prevalence of *Plasmodium* spp. in females ($\chi^2 = 6.49$; 1 d.f.; $P = 0.011$), but not in males ($\chi^2 = 0.30$; 1 d.f.; $P = 0.59$). Parents were caught with a door trap at the nestbox when nestlings were 14 days old. A drop of blood was taken from the brachial vein and a thin smear produced on a glass slide. Slides were air-dried, fixed with absolute methanol and stained with Giemsa stain. Each slide was screened with a light microscope under oil immersion for 10 min, and *Plasmodium* spp. scored as present or absent. The experiments were performed in a stable breeding population of great tits around the campus of the University of Lausanne, in spring 1995.



Bird species diversity

SIR — Graham Martin's report¹ of the recent symposium on avian systematics (*Avian Taxonomy from Linnaeus to DNA*; London, 23 March 1996) covered several important aspects of the debate over two different species concepts — the biological and phylogenetic — and how use of the latter might increase the number of bird species recognized. But he gave a somewhat partial view, and ornithologists do not need to "start saving for the double-sized revised editions".

In particular, Martin suggested that new methods of molecular analysis, such as geographical studies of mitochondrial DNA (mtDNA), would lead to a doubling of the currently recognized species diversity of birds. The figure Martin quoted of 20,000 species of birds, rather than the currently recognized ~9,000, did indeed come from my presentation, based in part on unpublished work by G. F. Barrowclough, J. Cracraft and myself. However, contrary to Martin's implication, that study involved neither molecular data nor minute morphological differences. It was based on re-evaluation of already known, broad patterns of morphological variation (as I indicated in my talk). Hence, the data to re-evaluate bird species diversity under alternative species concepts, such as a phylogenetic species concept², already exist. Many field guides already figure the distinctive subspecies that would probably be considered species under the phylogenetic species concept, and many bird watchers can already distinguish these forms. Thus, field guides would not double in size.

Two other points merit mention. First, it is likely that the number of phylogenetic species in the world will be much less than the number of already recognized subspecies — so the phylogenetic species concept will probably not greatly increase the number of names needed. Indeed, probably only 50% of recognized subspecies would qualify as phylogenetic species (our unpublished data, mentioned above). Second, there are very few studies in which several genes provide data for subdivisions of species that are not already apparent from morphological studies. One study that Martin cites makes this clear³. The bird pictured in his meeting report, the three-toed woodpecker (*Picoides tridactylus*), was the only species of 13 tested which showed major mtDNA distinctness without well-marked phenotypic variation (although the samples compared with

mtDNA were from different, previously recognized subspecies)⁴. This hardly fore-shadows a massive increase in species from molecular systematic studies. Thus, amateur and professional fieldworkers alike will not need DNA labs for field identification.

Molecular analyses are giving a refined assessment of patterns of variation within currently recognized biological species of birds. It is possible that there will be a 'class' of species defined by DNA markers which are difficult to distinguish in the field, but current evidence suggests this will not be an especially large category. There is indeed a serious debate about species

concepts, involving whether to give primary emphasis to the process of mate choice⁵ or evolutionary patterns^{6,7}, corresponding to the biological species concept and a phylogenetic species concept. Although more species of birds are likely to be recognized with both molecular data and the phylogenetic species concept, it will not make field identification much harder than it is at present, and will lead to an improved understanding of the biology of birds.

Robert M. Zink

Bell Museum,
University of Minnesota,
St Paul, Minnesota 55108, USA

Skin grafts and cheetahs

SIR — Since the publication of the landmark study by O'Brien *et al.*¹ on the lack of genetic variation in cheetahs, a flurry of reports have questioned this work^{2–6}. This scepticism is due, in part, to the acceptance of reciprocal skin grafts between unrelated cheetahs reported by O'Brien *et al.*, a phenomenon not previously observed in wild mammals. Such allogeneic graft acceptance suggests unusual monomorphism at the major histocompatibility complex (MHC), a group of loci responsible for many immune functions. These skin-graft results, a cornerstone of O'Brien *et al.*'s hypothesis on genetic vulnerability, have been criticized for problematic methodology and inconsistencies³. Recently, in *Nature*, May⁶ and Laurenson *et al.*⁷ characterized these trials as "dodgy" and "suspicious", respectively, citing concerns first raised by Caughley³. Surprisingly, no attempts have been made to repeat this test on other wild species³.

We performed skin-graft experiments in *Thomomys bottae*, the pocket gopher. The populations of this species together span the range of mean heterozygosities (H) found in all mammals ($H = 0.01–0.24$)⁸. We trapped individuals from two low-variation populations (Patricks J and Patricks F, $H \approx 0.02$) and one high-variation population (Hastings, $H \approx 0.16$), and performed reciprocal skin grafts on within-population pairs of animals. Each animal received one allograft in addition to an autograft as a control. We monitored grafts for 3 months for signs of rejection. Accepted allografts became indistinguishable from autografts typically within 2 weeks after surgery. The few technical failures caused through graft injury were excluded from our analyses.

Multi-locus DNA fingerprinting analyses were performed on all individuals and band sharing was assessed⁹. Hastings animals showed low band sharing (0.55, $n = 16$), whereas the Patricks J and Patricks F animals showed extreme sharing of bands within populations (0.97, $n = 21$; 0.98, $n = 7$, respectively) but low band sharing between populations (0.31).

We used a subset of these animals for skin grafting. Essentially all Patricks J ($n = 14$) and Patricks F ($n = 6$) individuals accepted within-population allografts, whereas Hastings animals ($n = 12$) rejected all such grafts. The difference in acceptance between high- and low-variation populations was highly significant ($\chi^2 = 30$, $P < 0.0001$). To test immunocompetence in the low-variation populations, we gave allografts from Hastings animals to each of three Patricks J and two Patricks F animals that accepted within-population allografts; these grafts were rejected.

Skin graft acceptance is an indicator of genetic near-identity, at least at the MHC. Our results indicate that individuals from populations with low levels of genetic variation can have similar MHC genotypes, and we believe that the earlier cheetah results¹, because of their concordance with our work, were indeed real phenomena. A consequence of such extreme homozygosity could be a loss of fitness caused by a reduction in immune-system pliancy^{1,10}. Patricks J and Patricks F gophers, like many endangered species, appear to be highly homozygous, but are not immunocompromised and manage to persist in the wild. Such populations, however, assuming a concordance between MHC variability and disease susceptibility, may be equally vulnerable (or resistant) to particular pathogens.

M. A. Sanjayan, Kevin Crooks

Department of Biology,
Sinsheimer Laboratories,
University of California,
Santa Cruz, California 95064, USA

1. O'Brien, S. J. *et al. Science* **227**, 1428–1434 (1985).

2. Caro, T. M. & Laurenson, M. K. *Science* **263**, 485–486 (1994).

3. Caughley, G. J. *Anim. Ecol.* **63**, 215–244 (1994).

4. Merola, M. *Conserv. Biol.* **8**, 961–971 (1994).

5. O'Brien, S. J. *Conserv. Biol.* **8**, 1153–1155 (1994).

6. May, R. M. *Nature* **374**, 309–310 (1995).

7. Laurenson, M. K. *et al. Nature* **377**, 392 (1995).

8. Patton, J. L. & Smith, M. F. *The Evolutionary Dynamics of the Pocket Gopher Thomomys bottae, with Emphasis on California Populations* (Univ. California Press, Berkeley, 1990).

9. Zimmerman, R. C. *SIM 1.1 Prog.* (Univ. California, Los Angeles, 1993).

10. Thorne, T. E. & Williams, E. S. *Conserv. Biol.* **2**, 66–74 (1988).

1. Martin, G. *Nature* **380**, 666–667 (1996).
2. Cracraft, J. *Curr. Ornith.* **1**, 159–187 (1983).
3. Zink, R. M. *Evolution* **48**, 96–111 (1994).
4. Zink, R. M. *et al. Condor* **97**, 639–649 (1995).
5. May, R. M. *Proc. Calif. Acad. Sci.* **48**, 131–140 (1993).
6. Graybeal, A. *Syst. Biol.* **44**, 237–250 (1995).
7. Davis, J. I. *Syst. Bot.* **20**, 555–559 (1995).

Recycling history

Sydney Brenner

Making PCR: A Story of Biotechnology. By Paul Rabinow. University of Chicago Press: 1996. Pp. 190. \$22.50, £17.95.

A COMMON genre of writing about science seeks to sensationalize the scientific endeavour and put it on a par with other notable activities in our society such as leveraged buyouts and the corruption of politicians. Titles such as *Getting to the Last Base: The Race for the Human Genome* or *The Nobel Art of Prizefighting: A Bout with Viruses*, although invented, would not go amiss today when nothing can qualify for attention without its fair share of sleaze. The rise of biotechnology and the convergence of academia and the marketplace offer good opportunities not only for journalists but also for the serious student of the scientific life.

Paul Rabinow's *Making PCR* is a serious book that "focuses on the emergence of biotechnology, circa 1980, as a distinctive configuration of scientific, technical, cultural, social, economic, political and legal elements.... It examines the 'style of life' or form of 'life regulation' fashioned by the young scientists who chose to work in this new industry rather than pursue promising careers in the university world". Rabinow uses the example of PCR (the polymerase chain reaction) to discover how this group of scientists and others view themselves and their work. As a social anthropologist he is interested in how this departs from the traditional scientific value system and he thinks that the discrepancies between the views of those who work in science and those who study scientists are themselves an interesting form of study.

It is best first to give a brief introduction to PCR and some of the events that surrounded its invention and later applications. PCR allows the exponential amplification of selected segments of DNA. Two short oligonucleotides, corresponding to sequences flanking the segment, are annealed to denatured DNA and these primers are then extended using a polymerase. This cycle of denaturation, annealing and extension is repeated as many times as desired, the DNA doubling in each cycle. A thermostable enzyme *Taq* I polymerase, which survives the heating

required for the denaturation step, facilitates the automation of the process. The first paper on PCR was published in *Science* in 1985 by seven scientists from Cetus, a California biotechnology company, with the name K. Mullis exactly in the middle of the list. Kary Mullis is acknowledged as the inventor of PCR and received the Nobel prize for chemistry for this work in 1994. In 1989, *Science* published in its December issue the first of its "Molecule of the Year" articles, devoting it to PCR and *Taq* I polymerase. That year

'justness' of the award of the Nobel prize to Mullis alone. Most believe he exaggerated his own claims, and a strong feeling emerges, as it does through the whole book, that he never quite understood what he was doing and that nothing would have happened had not others worked hard to achieve its implementation.

When the *Science* article appeared, PCR was greeted as a revolution. It was not that at all; it was a tool, and another in the long list that allow us to interrogate nature directly. Mullis by now has had considerable experience in designing the history of PCR, and has defined PCR as a concept and not a technique. His description of his "'Eureka!' thing" was that he "put together elements that were already there"; indeed, the method of annealing short lengths of oligonucleotides to complementary strands was already known and used in sequencing, as was chain



"The Return of Jonah" by John Wehrle, a painting commissioned by Tom White and Kary Mullis for their laboratory at Cetus. From the cover of *Making PCR*.

had marked an explosive rise in the numbers of papers dealing with PCR and its applications. Mullis had already left Cetus in 1986, the patents covering PCR began to be issued in 1987 and, around the same time as the 1989 *Science* article, Du Pont challenged the PCR patent. The patents were upheld in February 1991 and, by the end of the year, Cetus had disappeared, with Hoffmann-La Roche acquiring PCR for \$300 million and Chiron the rest of Cetus's assets.

The book includes several interviews with all the main characters involved, so the development is traced not only through the eyes of the observer but also, as in "Rashomon", through the words of those who took part in the events. We have to be careful here because memories of the past can become transformed by events in the future.

The concluding chapter is most interesting. Here the interviews are especially illuminating because they concern the

extension. If the idea of cycling is what Mullis believes he invented, and if the idea contains the essential invention, then that had already been exactly described by Khorana. It is one thing to ask why others had not thought of PCR before Mullis, but not only had Khorana thought about it but he also wrote it down. An invention also requires an implementation, but it is clear that this was done by others. Perhaps it is the implementation that converts an idea into a concept, but I have heard it said that while an idea is not a patentable invention, writing a patent on it may make it one.

There is much in Mullis's description of his discovery that echoes the adolescent romantic view of the scientist as the individual genius, unknown and unrecognized, working away in his garage laboratory, a picture that has analogies with the Bohemian artist or poet in his garret. Rabinow explodes this notion and shows that although ideas may locate in

one brain at one time, getting the science out into the real world is a complex social process involving many people and institutions. Both the desirable and the undesirable effects of science are produced by this process, and those who wish to regulate it cannot do so by going to the fountain head and stopping research.

Actually, PCR would make a wonderful subject for the more sensational science writer. Very few of the dramatis personae

come out as totally admirable and only one character comes out well in the author's view — and that, I surmise, is the subject of the dedication of the book ("Monsieur Le Blanc"). Aspiring journalists should take note that I have already trademarked the title: *Making it in Biotechnology: The Story of PCR*. □

Sydney Brenner is a fellow of King's College, Cambridge CB2 1ST, UK.

The voice of experiments

William Shea

Discipline and Experience: The Mathematical Way in the Scientific Revolution. By Peter Dear. University of Chicago Press: 1995. Pp. 290. \$60, £47.95 (hbk); \$24, £19.25 (pbk).

Scientific Practice: Theories and Stories of Doing Physics. By Jed Z. Buchwald. University of Chicago Press: 1995. Pp. 398. \$65, £51.95 (hbk); \$24.95, £19.25 (pbk).

The Scientific Voice. By Scott L. Montgomery. Guilford: 1996. Pp. 459. \$44.50 (hbk); \$19.95 (pbk).

As the twentieth century comes to a close, politicians and administrators pat applied science on the back and turn a cold shoulder on pure science, chided for wasting its time on understanding the world when it should be finding ways of transforming it to create new jobs and reduce unemployment. Chameleon-like, historians of science have taken the hue of their political paymasters and have begun talking 'more seriously' about the practical side of scientific research. Without denying outright that theory is relevant, historians have rediscovered the crucial role of experiments and the essential contribution of technology. While ideas are undoubtedly interesting, they are said to remain sterile until validated by experiments. Better still, experiments give rise to new ideas and channel them into useful and financially rewarding paths. This new trend may sometimes appear too fashionable to be the outcome of a serious reappraisal of the history of science, but it is nonetheless a welcome corrective to the unilateral emphasis on philosophical notions such as the rise of paradigms or their overthrow in moments of revolutionary fervour.

In *Discipline and Experience*, Peter Dear examines the Scientific Revolution in the light of this growing interest in experimentation, and he offers a reconstruction of the evolution of mathematical physics where decisive experiments, such as those of Torricelli, Boyle and Pascal, are thrust to the fore. Dear also discusses the relevance of methodological problems and he provides a succinct but comprehensive sur-

vey of the work of Jesuit scientists. It is interesting to find that the Jesuits carried out a lively debate on the nature of free fall. Giovanni Battista Riccioli, the most famous astronomer of his day, criticized the experimental rigour of his fellow-Jesuit, Niccolò Cabeo, and offered the results of his own observations. "Besides wooden balls", he writes, "we released stones of diverse weights from the tower of our chapel of the Society of Jesus, and let them fall once on a bronze basin, and at another time on a wooden board, so that from the different sounds I could distinguish better which one reached the ground faster". The heavier stones clearly arrived before the lighter one, and Riccioli main-



Naming of parts: Johannes Hevelius's map of the Moon, 1647.

tained that the theory that the speed of falling bodies is independent of weight was a hasty generalization refuted by experience.

The title of the collection of 12 essays edited by Jed Z. Buchwald, *Scientific Practice: Theories and Stories of Doing Physics*, would seem to indicate a return to theoretical concerns, but this is not the case. What the authors do, from a wide variety of vantage points, is examine small-scale experiments in order to gain a better insight into the relationship between theory and experiment. Five authors, Peter

Galison, Andrew Pickering, Hans Radder, Brian Baigrie and Yves Gingras, explore general issues such as objectivity, the reproducibility of results and the cultural constraints that influence the intentions of researchers. Gingras offers a particularly spirited plea for listening to the voice of experiments over and above the chatter of sociologists. Five other authors, Buchwald, Giora Hon, Margaret Morison, Simon Schaffer and Andrew Warwick, present the results of close examinations of specific instances of experimental research in the late nineteenth and early twentieth centuries. These include Hertz's pioneering investigation of cathode rays, the experiments that Kaufmann adduced as evidence against both Lorentz's theory of the electron and Einstein's theory of relativity, the analysis of sunspot spectra, and the attempts of Joseph Larmor and Frederick Trouton to generate unlimited amounts of energy from the contraction of matter as the Earth moves through the ether.

Historians and philosophers of science will find much to admire, ponder and discuss in the book by Dear and the essays edited by Buchwald. Scientists with a broad interest but no particular training in history or philosophy will find a more congenial writer in Scott L. Montgomery, the author of *The Scientific Voice*. A geologist with a passion for the intellectual tools of his trade, Montgomery describes himself as an amateur, but he deserves to be called a virtuoso in the sense in which this word was applied to members of the Royal Society in the seventeenth century. Over the past decade he has published a number of thought-provoking essays that have been revised for inclusion in this book. Montgomery is fascinated with scientific language, how it is born, how it grows, how it is transformed and how it is sometimes stultified. His interest is not an academically driven exercise but springs from an infectious desire to know how science works.

General reflections on the jargon of hard science in the first chapter lead to the soft but prickly discourse of psychoanalysis in the concluding essay. The four other chapters deal with the history of scientific writing, the military metaphor and the fight against disease, Japanese science and the politics of translation, and the many faces and names of the Moon. This last essay, entitled "Seeing and Naming the Skies - The Case of the Moon", is an original contribution to the relationship between mapping and per-

ceiving. The first naturalistic drawing of the Moon in Western culture by Jan Van Eyck dates from around 1420. Why so late? asks Montgomery, who proceeds to show that it was only when astronomers started looking at the Moon as another Earth that they began seeing it for what it really is. Whereas Thomas Harriot, the first person to produce an image of the Moon on the basis of telescopic observations in 1609, saw only "a strange spottedness all over", Galileo saw mountains and valleys and a huge circular crater that he compared in size and shape to Bohemia.

Galileo looked at the Moon with the eye of a cartographer. His map was a kind of visual experiment, a way of 'demonstrating' that the bright spots on the Moon's surface were tips of mountains, and the curiously shaped shadows were deep valleys. The science of optics and the art of map-making could achieve little alone. It was their combination that opened a new era of investigation. Mapping the Moon became a rigorous science as well as a political act; astronomers and their patrons tried to make the Moon their own by naming its features. Modern scientists are less prone to plant flags on virgin soil but they can be very keen about patents.

Montgomery offers a survey of the field of scientific illustrations and the ways in which they have been read, not only in terms of their iconography and pictorial effect, but also as embodiments of changing ideologies and shifting epistemologies. This 'rhetoric' of imagery is fraught with internal contradictions, competing messages and the chance that the visual representation may be perceived in new ways not intended by its maker.

Images in scientific journals are always undergoing some kind of extension and innovation, and there is often a tension between text and illustration. Two universes of readerly experience are sometimes laid side by side. One is weighed down by jargon and flatness in writing, while the other grows more colourful and appealing year by year. The visual material is getting so good that it tends to be there for aesthetic rather than illustrative purposes. No word is innocent. The same can be said of illustrations.

Montgomery combines sound scholarship with a sense of humour, which occasionally becomes a sense of urgency when experiments with language get out of hand and acquire a canonical status they do not deserve. This book will be read with enjoyment by all those who suspect that the truths of the present are often the metaphors of the past. □

William Shea is at the Institut d'Histoire des Sciences, Université Louis Pasteur de Strasbourg, 7 rue de l'Université, 67000 Strasbourg, France.

Dinosaurs for grown-ups

Alan Charig

The Evolution and Extinction of the Dinosaurs. By David E. Fastovsky and David B. Weishampel. Cambridge University Press: 1996. Pp. 460. £29.95, \$44.95.

UNDERGRADUATE courses on dinosaurology are now common in North America. This book is intended mainly as a textbook for these courses, but it serves also to bring specialists up to date. It provides a readable account of nearly every aspect of dinosaurs. There is, however, little descriptive anatomy; that subject is already well covered by *The Dinosauria* edited by D. B. Weishampel, P. Dodson and H. Osmólska (University of California Press, 1990).

Most aspects of dinosaurs are beset with controversy (for example, origins, bipedality, phylogeny, endothermy, origin of birds and extinction); this is what makes them so fascinating. An excellent feature of the new book is its airing of alternative views, usually explained without bias, although the authors themselves often plump for one or the other. Unfortunately, however, these views are sometimes put forward without attribution to their original proposers or without bibliographical reference. Similarly, other authors' figures are sometimes copied without acknowledgement.

The book is divided into four parts. The five chapters in the first part cover taphonomy, collecting and preparation; stratigraphy and chronology, plate tectonics and palaeoclimatology; hierarchy and evolution; the interrelationships of all life forms (with the emphasis on the chordates); and the origin of the dinosaurs. The five chapters in the second part deal with the various groups of Ornithischia (unusual, to do them first); each chapter is organized into a history of discoveries, definition and diagnosis, diversity and phylogeny and, finally, palaeoecology and palaeobiology. The third part (three chapters) is concerned, in rather less detail, with the Saurischia and the origin of birds, whereas the fourth part (four chapters) covers endothermy, distribution in space and time, and extinction. The text includes 20 boxes and several footnotes on incidental topics. "Important readings" are given at the end of each chapter, but there are not

nearly enough. An excellent glossary at the back is followed by three separate indexes: for subjects, genera and authors. A few anecdotes, poems and cartoons provide light relief; and the writing style, although a little too informal for me, is thoughtful and interesting and will probably please student readers.

The black-and-white illustrations include photographs, drawings (somewhat stylized), graphs, charts, tables and maps. Fourteen colour paintings by Brian Regal are placed together near the end; they are no doubt accurate, but again very stylized and (to my eye) devoid of charm.

The authors employ a purely cladistic approach to phylogeny reconstruction and classification, without hierarchical ranks in the latter; they explain cladistics using non-biological analogies, classifying motor vehicles and wrist-watches accordingly. But there is virtually no mention of ancillary methods of phylogeny reconstruction, such as stratigraphical chronology, embryology and ontogeny and, best of all, transformation series; nor is there any suggestion that synapomorphies (homologous characters) should be chosen critically. Like all cladists, the authors eschew stem-groups and other paraphyletic taxa. But as all these cannot be avoided in discussion, they are disguised in various ways: *Thecodontia* as *basal archosaurs* or simply '*Thecodontia*', *Prosauropoda* as *basal sauropodomorphs*, and *dinosaurs* in the usual sense as *non-avian dinosaurs* (although, in fact, the



TERTIARY amphibians, as depicted in *Life Before Man* by Z. V. Špinar. This classic book, first published in 1972, inspired a whole generation of palaeontologists with its evocative sequence of colour illustrations by Zdeněk Burian. Many new ones appear in the revised and updated paperback edition, which now covers recent discoveries of fossil fishes, amphibians and reptiles and the latest theories about the origins and early development of mammals and humans. Thames and Hudson, £9.95.

authors themselves then announce that they are going to abbreviate the latter to *dinosaurs*!).

To my mind, this is absurd. Classifications have purposes other than the indication of direct ancestor–descendant relationships. What is wrong with naming groups, especially stem-groups, that we all have to use all the time? Moreover, I am genetically much closer to my first cousin than to my descendant a thousand generations down the line, and the logical conclusion of a Russian-doll type of classification is that I am a blue-green alga. But this is not the place for an orgy of cladist-bashing; the authors are entitled to do it their own way, at present so fashionable among the younger palaeontologists of North America.

What is deplorable, however, is the redefinition of existing words, words in frequent use, to mean something utterly different. Even more confusing is the indiscriminate use of such words in both senses. Thus 'monophyletic' is used to describe the Sauropoda as "evolved from a single ancestor" (Ernst Haeckel's original definition, correct etymologically and in every other way; antonym 'polyphyletic'). Four lines later, however, the term is applied to the Prosauropoda in a context where it can only mean "including all its own descendants, without end" (as redefined much later by Willi Hennig; antonym 'paraphyletic'). Meanwhile, the authors have every right, since they so wish, to split the Amniota vertically into Reptilia (excluding mammal-like reptiles and including birds) and Synapsida (including mammals); but why misuse those familiar names — especially as Edwin Goodrich had invented new names for the same groupings, Sauropsida and Therapsida respectively, some 80 years ago?

I have some further brief comments. Most important, there is no indication of how a dinosaur is defined, beyond a statement that all dinosaurs share a host of derived characters; most of those are wrong, and a few others that are correct have been omitted. It is also stated that the earliest dinosaurs were "irrevocably bipedal" (which sits strangely with the following statement that all quadrupedal dinosaurs are secondary reversions); there is not a shred of evidence for any of this. Replacement of the term "fully improved" limb posture by "fully erect" is misleading; students often confuse "erect" with "bipedal" (which, of course, it isn't) and the potential misconception is reinforced by the figure on page 83. The origin of the dinosaurs, no later than the Carnian, could hardly have been induced by the much later alleged extinctions in the Late Triassic; their supposed derivation from the lagosuchids is also unsupported by any real evidence. There is no mention of dinosaur finds before 1818, such as Reverend Dr

Robert Plot's *Megalosaurus* femur from Oxfordshire (1676) or William Smith's *Iguanodon* tibia from East Sussex (1809). With regard to predator–prey ratios, the number of skeletons available for fossilization is in inverse proportion to the unknown length of the average life span (is this what the authors rather obscurely refer to as "turnover"?); the number can therefore tell us nothing about the size of the living population, which fact alone renders Robert Bakker's method useless as an indicator of endothermy and makes all the other uncertainties superfluous. Surprisingly, the authors favour a catastrophe theory for the extinction of the dinosaurs; their attempted refutation of the principal

objection to all catastrophe theories, namely that many major groups of animals and plants survived the supposed extinction event virtually unscathed, is far from convincing.

Despite my severe criticisms on certain specific points, I still think that this is the best introductory textbook for students. It is also extremely useful to the specialist; I only wish it were better referenced. Even so, I would not be without it. □

Alan Charig is an active retired worker (former chief curator of fossil reptiles and birds) in the Department of Palaeontology, The Natural History Museum, London SW7 5BD, UK.

Global movers and shakers

Jean-Pierre Burg

Tectonics. By Eldridge M. Moores and Robert J. Twiss. W. H. Freeman: 1995. Pp. 415. \$59.95, £29.95.

In the seventeenth century, naturalists began to see structures (*structurae* in Latin) in rocks. These *structurae* were considered to be primary features, acquired while rock layers were being deposited, but the term was quickly extended to deformation features that nearly 200 years later were established to be tectonic structures (the ancient Greek *τεχτονική* means art of building). Ever since, the academic difference between structural geology and tectonics in the study of the 'architecture' of the Earth has been blurred. Although structural geology does not exclude a geological object because of its size, tectonics generally deals with the larger features. The title *Tectonics* stems from these intricate semantics, and duplications of text and figures found in the authors' companion *Structural Geology* (W. H. Freeman, 1992; for a review see *Nature* 361, 416: 1993) do not help to clarify the distinction.

Structural geologists overcame a large psychological barrier with their discovery of evidence of horizontal forces and movements. Using the distribution of mountain ranges to define belts of continental shortening, and fold orientations to determine the force directions, they imagined slow yet small motions between continental blocks. A few visionary geologists embraced the idea of continental drift developed by Alfred Wegener in 1915. But they were unable to prove the existence of plate movements because they needed to explore the oceans, not the continents, to find convincing arguments.

E. M. Moores and R. J. Twiss reflect on the setbacks that have paved the path to success of plate tectonics and comprehensively document how the return of research to the sea, successfully done by geophysi-

cists in the 1960s, was vital for teaching us about the versatility of the Earth's outer shell. They make this point before these historical considerations, by describing most of the geophysical techniques and discoveries that brought about a conceptual revolution in the Earth sciences. The text avoids jargon, and technical terms are defined in footnotes and boxes. The foundations of plate tectonics are solidly laid out through an interpretation of the recent, large-scale features of the Earth in terms of relative movements (the kinematics) and forces (the dynamics). Plates collide and compression dominates at convergent plate boundaries; plates are pulled apart and tension dominates at divergent plate boundaries; and plates slide horizontally past each other and shearing dominates at transform plate boundaries. Each of these three tectonic settings and the resulting characteristic structures are accurately reviewed.

The authors are less successful in their treatment of the case studies of orogenic belts. For example, in the past ten years structural geologists have shown that thickened continental crust may deform in response to forces other than those due to plate motions. Regrettably, this is not considered by the authors. Nor do they describe a few key techniques for understanding large-scale structures such as numerical modelling. Had they avoided the duplications mentioned above, the authors would have had the space to cover the latest tectonics research and so produce a stronger reference for undergraduate and graduate students as well as scholars. Nevertheless, every reader will grasp and appreciate the knowledge presented in this clear, generously illustrated companion to *Structural Geology*. □

Jean-Pierre Burg is at the Geologisches Institut, ETH-Zentrum, Sonneggstrasse 5, CH-8092 Zürich, Switzerland.

Molecular chaperones in cellular protein folding

F. Ulrich Hartl

The folding of many newly synthesized proteins in the cell depends on a set of conserved proteins known as molecular chaperones. These prevent the formation of misfolded protein structures, both under normal conditions and when cells are exposed to stresses such as high temperature. Significant progress has been made in the understanding of the ATP-dependent mechanisms used by the Hsp70 and chaperonin families of molecular chaperones, which can cooperate to assist in folding new polypeptide chains.

PROTEIN folding is the process by which the linear information contained in the amino-acid sequence of a polypeptide gives rise to the well-defined three-dimensional conformation of the functional protein. How this is accomplished constitutes a central problem in biology. Because unfolded proteins can reach their native state spontaneously *in vitro*¹, it was assumed that the folding (acquisition of tertiary structure) and assembly (formation of protein oligomers) of newly synthesized polypeptides *in vivo* also occur essentially uncatalysed and without the input of metabolic energy. This long-held view has been revised in recent years owing to the discovery that in the cell the correct folding of many proteins depends on the function of a pre-existing protein machinery—the molecular chaperones^{2–5}. Here I summarize our present understanding of the structure and function of these components. The main focus will be on the ATP-dependent mechanisms used by the heat-shock protein 70 (Hsp70) and the chaperonin families (Table 1), which fulfil distinct roles in protein folding: the Hsp70s and their partner proteins can stabilize translating and newly synthesized polypeptides until all segments of the chain necessary for folding are available; the chaperonins are large cylindrical complexes that promote protein folding in the sequestered environment of their central cavity.

Molecular chaperones were originally defined as a group of unrelated classes of proteins that mediate the correct assembly of other proteins, but are not themselves components of the final functional structures². They occur ubiquitously and many of them are classified as stress proteins, although they have essential functions under normal growth conditions. More recently, chaperones have been defined as proteins that bind to and stabilize an otherwise unstable conformer of another protein—and by controlled binding and release, facilitate its correct fate *in vivo*, be it folding, oligomeric assembly, transport to a particular subcellular compartment, or disposal by degradation⁵. Molecular chaperones do not contain steric information specifying correct folding; instead, they prevent incorrect interactions within and between non-native polypeptides, thus typically increasing the yield but not the rate of folding reactions. This distinguishes them from the so-called folding catalysts, protein disulphide isomerases and peptidyl–prolyl isomerases. These enzymes accelerate intrinsically slow steps in the folding of some proteins, namely the rearrangement of disulphide bonds in secretory proteins and the *cis*–*trans* isomerization of peptide bonds preceding proline residues, respectively (reviewed in refs 6,7).

Origins of the chaperone concept

Although the term ‘molecular chaperone’ was coined to describe the specialized function of the nuclear protein nucleoplasmin in promoting chromatin assembly⁸, our understanding of the general role of chaperones in protein folding emerged mainly from

research on the Hsp70s and the chaperonins. Based on their striking heat inducibility, it was proposed that the Hsp70s aid the repair or degradation of polypeptides that denature under stress^{9,10}. ‘Guilt by association’ indicated that Hsp70 has an equally important role under non-stress conditions in stabilizing the folding and assembly intermediates of newly synthesized polypeptides: the Hsp70 homologue in the lumen of the endoplasmic reticulum (ER), BiP, was found in a complex with unassembled immunoglobulin heavy chains¹¹ and other secretory proteins (reviewed in ref. 4). Similarly, cytosolic Hsp70 in yeast associates transiently with newly synthesized precursor polypeptides, stabilizing them for uptake into the mitochondrion and the ER^{12–14}. Mammalian Hsp70 was shown to interact with a large fraction of nascent polypeptide chains in the cytosol¹⁵, on the basis of its ability to bind hydrophobic peptide segments in an ATP-dependent manner¹⁶.

The functional analysis of the chaperonins provided the most compelling evidence that protein folding *in vivo* requires molecular chaperones. *Escherichia coli* cells with mutations in the gene coding for the chaperonin, GroEL, were shown to be defective in assembling bacteriophage particles^{17,18}. Curiously, it was the analysis of the assembly of ribulose biphosphate carboxylase oxygenase (Rubisco) in chloroplasts that clarified the significance of this finding. Before assembly, the Rubisco subunits form a complex with a large binding protein from which they dissociate in an ATP-dependent reaction^{19,20}. This binding protein was subsequently identified as the chloroplast homologue of GroEL and of the mitochondrial chaperonin, Hsp60 (refs 21,22). Indeed, purified GroEL allowed the reconstitution of correctly assembled bacterial Rubisco *in vitro*²³, and the analysis of a yeast mutant defective in Hsp60 revealed that the biogenesis of many proteins imported into mitochondria required the chaperonin²⁴. Finally, the mitochondrial system provided evidence that the chaperonins mediate the folding of monomeric proteins, rather than actively promoting assembly²⁵.

A puzzling conclusion from these independent lines of research was that cells contain at least two different chaperone systems with apparently overlapping functions. A coherent view emerged from the demonstration that the Hsp70 and chaperonin families can cooperate functionally *in vitro*²⁶. The Hsp70 homologue of *E. coli*, DnaK, and its partner chaperone DnaJ maintain an unfolded polypeptide in a soluble, folding-competent conformation, in which it can be transferred to GroEL for folding to the native state. There is increasing evidence that similar chaperone pathways operate in the cytosol and organelles of eukaryotic cells.

Folding *in vitro* versus folding *in vivo*

Significant progress has been made in recent years in understanding the kinetics and pathways of spontaneous refolding *in*

TABLE 1 Components of the Hsp70 and chaperonin systems in bacteria and eukaryotic cells

Hsp70 system	Bacteria	Compartment	Eukaryotic homologues	
			Yeast	Higher eukaryotes
Hsp70 ~70K ATPases, bind extended peptides, interact with DnaJ homologues and GrpE		Cytosol	SSA1-4; heat-inducible and constitutive forms	Hsp72: stress-inducible
		Mitochondria	SSB1,2: bind to polysomes	Hsc73: constitutive; binds to nascent chains
			SSC1: required for protein import and folding	mHsp70
	DnaK	ER lumen	Kar2/BiP: required for post-translational protein import and folding	BiP: abundant, binds unfolded ER proteins
Hsp40 ~40K proteins, bind unfolded proteins, interact with matching Hsp70s and stimulate ATP hydrolysis		Cytosol	YDJ1: interacts with Hsp70 SSA1; involved in protein transport across membranes	Hsp40: binds nascent polypeptides
		Mitochondria	SIS1: similarity to mammalian Hsp40	Hdj1: human Hsp40 homologue
	DnaJ	ER	MDJ1: required for folding of newly imported proteins	
			SCJ1 (lumen), Sec63 (membrane)	
GrpE ~20K nucleotide-exchange factor for Hsp70		Mitochondria	MGE1: interacts with Hsp70 SSC1	
	GrpE			
Chaperonin system				
GroEL family (Cpn60) Two rings of 7 ~60K subunits; ATPase activity, bind protein-folding intermediates, promote folding together with Hsp10 (Cpn10) cofactor		Mitochondria	Hsp60/Hsp10: required for folding of imported proteins	Hsp60/Hsp10
		Chloroplasts		Cpn60*: two homologous subunits (α and β); cooperates with cpn20† in folding Rubisco subunits
	GroEL/GroES			
TRiC‡ family Hetero-oligomeric, 2 rings of 8 ~55K subunits; ATPase activity, bind protein-folding intermediates, promote folding in eukaryotes and Archaea		Cytosol	Defects in genes encoding TRiC subunits result in cell-cycle mutants with defects in actin and tubulin assembly	TRiC particles purified, 8 related ~55K subunits with homology to TCP1

* Chloroplast Cpn60 is also known as Rubisco-subunit-binding protein.

† The subunits of Cpn60's cooperating factor Cpn20 consist of two fused Cpn10-homologous domains.

‡ TRiC is also known as CCT (chaperonin containing t-complex polypeptide).

vitro for small proteins (reviewed in refs 27,28) (Fig. 1a). Briefly: (1) upon removal from denaturant, the polypeptide chain collapses into a compact shape within milliseconds, driven by the tendency to bury hydrophobic residues. This coincides with the formation of secondary structure. (2) Native tertiary structure forms as these secondary elements interact with each other to assume defined positions in space. During this process, which usually takes seconds to minutes, the polypeptide may pass through a series of kinetically defined intermediates. (3) The major rate-limiting transition state in folding lies close to the fully native state. In the case of multidomain proteins, individual domains fold first, followed by their association to folded monomers. Monomers may then associate to form oligomers.

It seems highly unlikely that two fundamentally different mechanisms exist *in vitro* and *in vivo* to explain a reaction as complex as protein folding. Yet folding *in vivo* depends on a cellular machinery that prevents protein aggregation. To understand this requirement, we have to consider the conditions in which *de novo* folding differs from *in vitro* refolding: (1) the cell cytosol is a highly crowded macromolecular environment; and (2) folding in the cell must be accomplished in the context of the vectorial synthesis of polypeptide chains on ribosomes.

(1) **Molecular crowding and protein aggregation.** Unfolded polypeptides, including compact intermediates that contain various amounts of native-like secondary structure but lack a stable hydrophobic core ('molten globules'), expose hydrophobic amino-acid side chains and therefore tend to aggregate^{5,27}

(Fig. 1). More advanced intermediates on the folding pathway can also give rise to aggregation²⁹. The extent of aggregation *in vitro* can often be controlled by lowering the protein concentration and temperature. In contrast, cellular conditions dictate that without the intervention of a specific machinery, aggregation would outcompete correct folding, at least for a significant fraction of newly synthesized polypeptides. The concentration of nascent (ribosome-bound) chains in the cytoplasm of *E. coli* is as high as ~35 μ M, assuming a uniform distribution³⁰. Their effective concentration will be significantly greater, however, because of molecular crowding and because translating ribosomes are organized in polysomes. Molecular crowding refers to the fact that a considerable fraction of the cellular volume is occupied by protein and other macromolecules at a total concentration of ~340 g per litre and is therefore not available to other macromolecules³¹. Crowding is predicted to cause an increase in macromolecular association constants over those in dilute solution by several orders of magnitude.

(2) **Folding and translation.** Stable folding requires the presence of at least a complete protein domain (usually ~100–200 amino-acid residues in length). As the C-terminal ~30 amino acids of a translating polypeptide are topologically restricted by the ribosome, nascent chains remain unfolded until an entire domain has emerged (Fig. 1b). Thus, actively growing cells must maintain a large population of aggregation-sensitive nascent chains in a folding-competent conformation. This is thought to be achieved by the co-translational binding of molecular chaperones,

including Hsp70 and DnaJ, to the nascent chain^{15,32–35}, but a protective function of the ribosome itself may also be important³⁶. In any case, a single-domain protein will have to fold post-translationally, after its release from the ribosome (Fig. 1b).

In principle, polypeptides consisting of two or more consecutive domains may fold cotranslationally and thus avoid incorrect interactions between domains during folding (Fig. 1c). At present, however, only scant evidence for cotranslational folding is available^{35,37,38}, including the finding that secretory proteins can form disulphide bonds between cysteine residues distant in sequence as they are extruded into the ER lumen (for example, see ref. 39). Cotranslational folding in the eukaryotic cytosol may require the protective function of the Hsp70 system until a protein domain is completed, and for certain proteins may then involve the cooperation of a chaperonin³⁵ (Fig. 5). The following sections will consider the detailed mechanisms of these two major chaperone systems.

The Hsp70 chaperone system

The Hsp70s are a family of highly conserved ATPases of relative molecular mass about 70,000 (M_r 70K) found in prokaryotes and in most of the compartments of eukaryotic cells (Table 1). They have essential roles in protein metabolism under both stress- and non-stress conditions, including functions in *de novo* protein folding and membrane translocation, the degradation of misfolded proteins as well as regulatory processes. This versatility of the Hsp70s results from their basic function, which is to bind and release hydrophobic segments of an unfolded polypeptide chain in an ATP-hydrolytic reaction cycle. Whereas binding results in the stabilization of the unfolded state, controlled release may allow progression along the folding pathway.

Important insight came from the finding that this binding/release cycle is modulated by other proteins of the Hsp70 system, most notably the DnaJs or Hsp40s. The central feature of this protein class is a conserved J-domain (named for *E. coli* DnaJ) that has been implicated in the interaction of Hsp40s with their respective Hsp70 partners. In DnaJ, this interacting domain is coupled to a polypeptide-binding domain, so that recruitment of the Hsp70, DnaK, provides the combined functions of two chaperones and ATP-dependent cycling. By joining the J-domain with a variety of protein moieties, cells use the Hsp70/Hsp40 system not only to protect a range of transiently aggregation-prone substrates, but also for specialized cellular functions^{40–42}. For example, the mammalian protein auxilin binds to the clathrin coats of endocytotic vesicles, and via its C-terminal J-domain recruits the constitutively expressed Hsp70 (Hsc70) for ATP-dependent clathrin uncoating⁴¹.

Structure of Hsp70 and DnaJ. Three-dimensional structures of these proteins are only known in part. The Hsp70s are structurally subdivided into an N-terminal ATPase domain of ~45K, which is followed by an ~18K portion containing the peptide-binding site, and a more variable ~10K segment (Fig. 2a). The ATPase domain has a fold identical to the globular G-actin monomer⁴³ and transmits ATP-dependent conformational changes to the C-terminal peptide-binding domain⁴⁴. The ATPase domain of *E. coli* DnaK interacts with GrpE (ref. 45), DnaK's 22K nucleotide-exchange factor⁴⁴. DnaJ may also contact the ATPase domain. An analysis by multidimensional NMR⁴⁶ indicated that the peptide-binding domain of Hsc70 (ref. 47) consists of two four-stranded antiparallel β -sheets and a single α -helix. Thus it does not show the previously proposed structural similarity to the peptide-binding region of the human major histocompatibility complex (MHC) class I molecule.

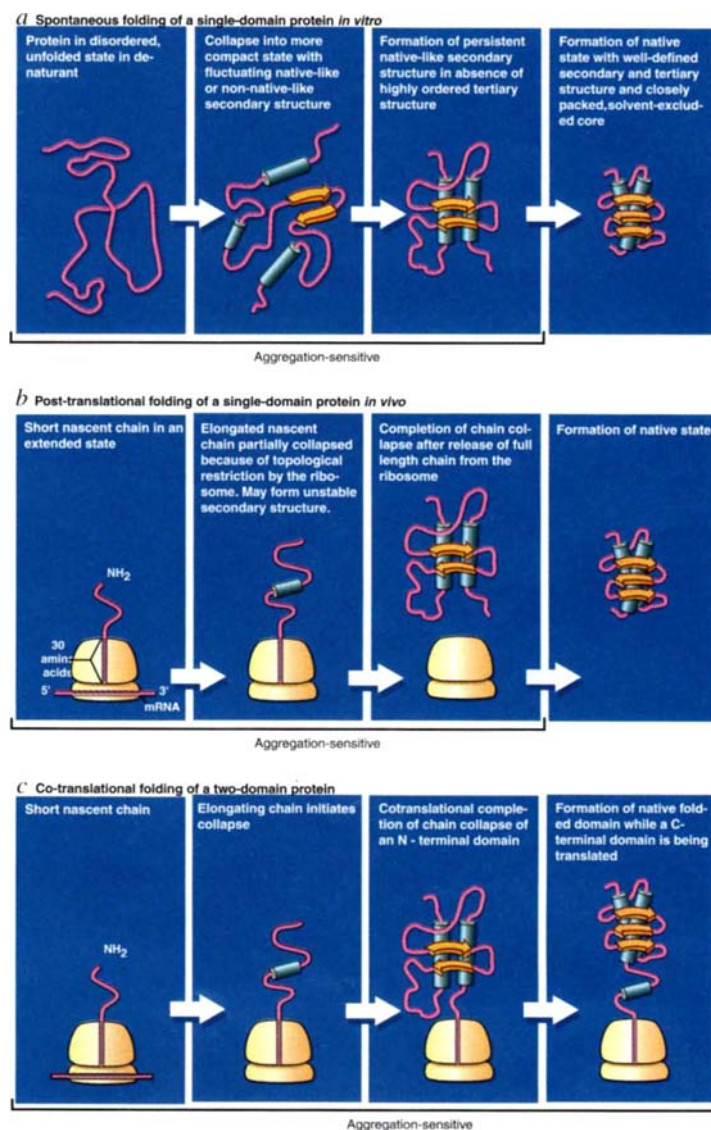
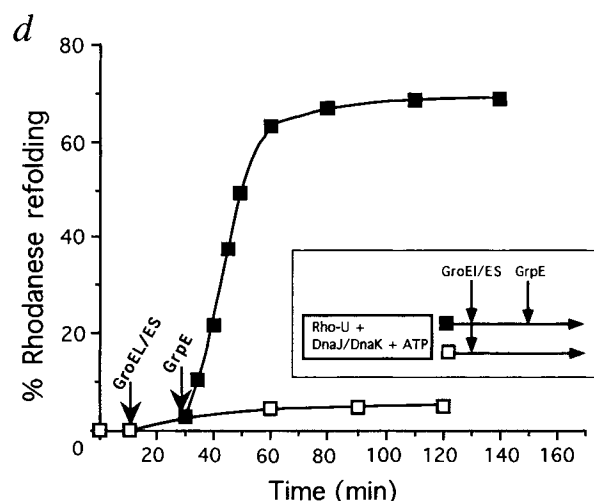
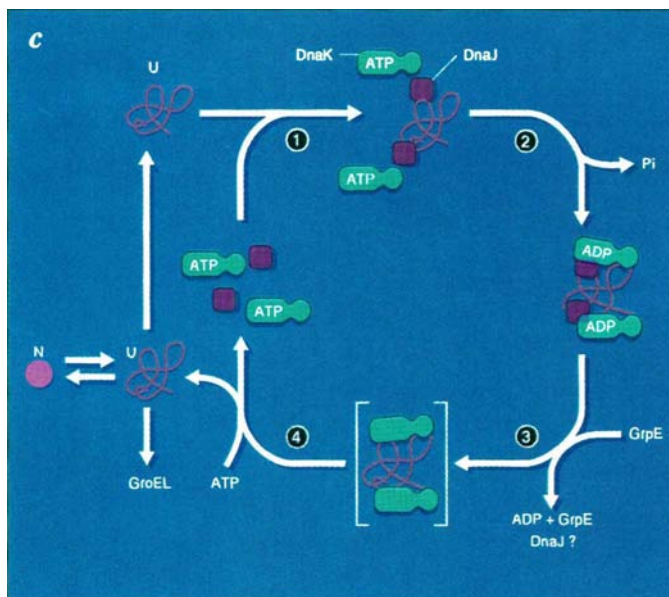
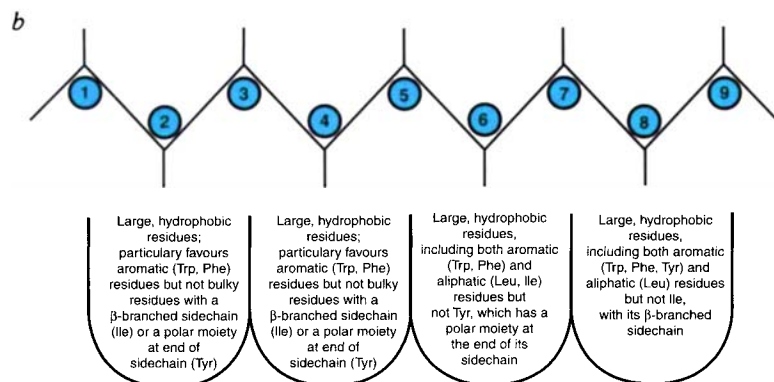
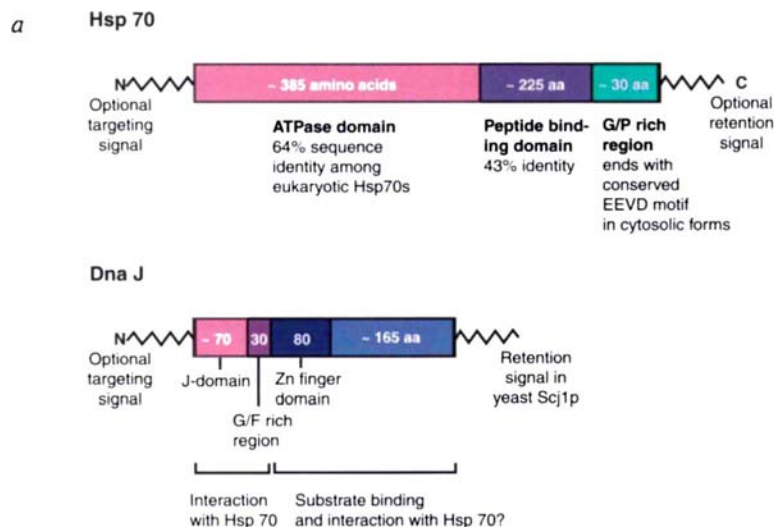


FIG. 1 Possible steps in protein folding shown for a hypothetical protein upon dilution from denaturant *in vitro* (a) and for translating polypeptides *in vivo* (b, c). Part a is adapted from ref. 153. Cylinders correspond to α -helices and arrows to β -sheets. Note that the synthesis of an average 40K polypeptide takes about 15 s in *E. coli* and 2–3 min in the eukaryotic cytosol.

Proteins of the Hsp40 family typically contain one or more blocks of sequence homology to *E. coli* DnaJ, a 41K protein with a clear domain structure (Fig. 2a). The N-terminal 70-amino-acid J-domain is present in all Hsp40s. The NMR-based structure of this domain describes a scaffolding of four α -helices, at the tip of which lies the conserved tripeptide His-Pro-Asp in a solvent-accessible loop region^{48,49}. Substitutions within this tripeptide abrogate the interaction with DnaK⁵⁰. Subtle structural differences between the J-domains of different Hsp40s seem to be critical in mediating their interaction with specific Hsp70 partners⁴². The J-domain is followed by a glycine-phenylalanine-rich region, a cysteine-rich central domain not present in all Hsp40s, and a C-terminal stretch found in most homologues (Fig. 2a). The central region contains two zinc atoms, each coordinated to four cysteines as in the zinc-fingers of DNA-binding proteins⁵¹. In DnaJ, this region and part of the C-terminal domain apparently participate in binding unfolded polypeptide.



Peptide binding. The various Hsp70s all have binding specificity for hydrophobic peptides, but their preference for certain amino-acid patterns may vary from compartment to compartment^{16,52-54}. The ER homologue, BiP, binds with micromolar affinity to 7–8-residue peptides^{16,52} corresponding to segments of a polypeptide chain that would be exposed in the unfolded state but are buried within the hydrophobic core of the folded protein. This explains the ability of Hsp70 to interact with a wide range of unfolded polypeptides while ignoring the native states of these proteins.

More than one Hsp70 may interact with an unfolded polypeptide. The peptides are bound in an extended conformation^{26,55,56} and a preferred pattern of certain aliphatic and aromatic residues at alternating positions in BiP-bound peptides has been described⁵² (Fig. 2b).

The binding specificities of Hsp40 and Hsp70 differ considerably: only Hsp70 interacts efficiently with α -lactalbumin which is in an extended conformation²⁶. On the other hand, *E. coli* DnaJ but not DnaK is efficient in preventing the aggregation of proteins

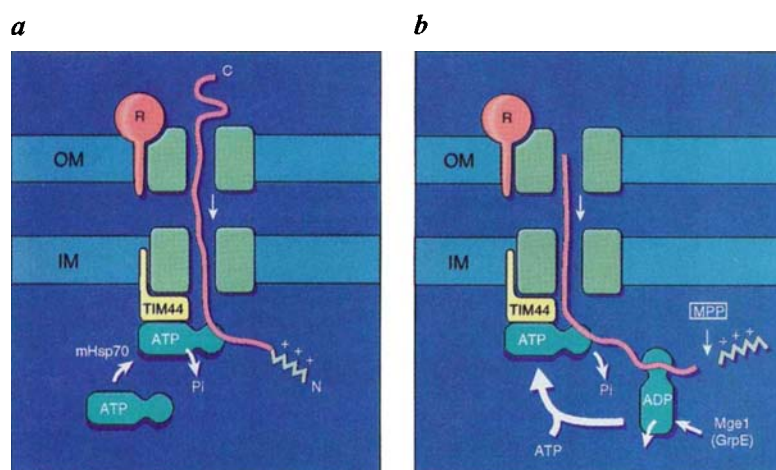


FIG. 3 Role of mitochondrial Hsp70 (mHsp70) and the GrpE homologue Mge1 in protein translocation. *a*, ATP-bound mHsp70 accepts a polypeptide segment from Tim44 (previously known as Mim44). *b*, mHsp70 binds polypeptide tightly as ATP is hydrolysed and translocation proceeds. The ADP form of mHsp70 is recycled by Mge1. R, receptor for precursor protein; OM, outer, and IM, inner mitochondrial membrane; MPP, mitochondrial processing peptidase (adapted from ref. 73).

upon dilution from denaturant^{26,57,58}. DnaJ can interact efficiently with the N-terminal segments of nascent polypeptides^{33,34} and combining the binding properties of DnaJ with those of DnaK appears to be essential for the function of this chaperone system in protein folding.

The Hsp70 reaction cycle in protein folding. The ATP-dependent cycle of peptide binding and release is best understood for the bacterial Hsp70 homologue DnaK and its cofactors DnaJ and GrpE. The ATP-bound form of DnaK binds and releases peptide rapidly. In contrast, the ADP-bound form has both slow on- and off-rates for peptide^{55,59}. DnaJ and GrpE, and peptide binding itself, accelerate the interconversion between these two states^{58,60}. An ATP hydrolytic reaction cycle has been proposed that parallels the regulation of certain GTP-binding proteins (Fig. 2c). (1) The chaperone DnaJ interacts first with an unfolded polypeptide, such as a translating chain, and then targets it to DnaK. (2) DnaK binds to the polypeptide in its ATP-bound state. The interaction with DnaJ then stimulates the hydrolysis of ATP by DnaK and stabilizes its ADP-bound state. This results in the formation of a stable ternary complex of unfolded substrate polypeptide, DnaJ and DnaK. (3) The nucleotide-exchange factor GrpE promotes the release of ADP from DnaK, which is rate-limiting in the cycle. ADP dissociation may destabilize the ternary complex and may cause the release of DnaJ. (4) Substrate dissociates from DnaK upon subsequent ATP binding (not hydrolysis) to DnaK^{59,61}. It then has the option of either folding, rebinding to DnaJ and DnaK, or being transferred to another chaperone system, such as the chaperonin GroEL, for final folding^{26,58} (Fig. 2c).

In each release event from DnaJ/DnaK, a fraction of the bound polypeptide may fold, as shown for firefly luciferase *in vitro*⁵⁸. The remainder, kinetically trapped as an unproductive intermediate, will rebind. The fraction of protein that folds upon release from DnaJ/DnaK is very small for mitochondrial rhodanese. Efficient folding of this protein depends on its GrpE-catalysed transfer from DnaJ/DnaK to GroEL²⁶ (Fig. 2d). Significantly, DnaJ/DnaK releases its substrate protein in an unfolded conformation^{58,62}, in contrast to the chaperonins (see below).

Although the prokaryotic and eukaryotic Hsp70 systems have similar functional properties, a GrpE-like nucleotide-exchange factor has not been found in the eukaryotic cytosol. It may be dispensable because the rate-limiting step in the ATPase cycle of eukaryotic Hsp70 is not the dissociation of ADP but rather the hydrolysis of ATP itself^{63,64}. In this system ADP dissociation is even subject to negative regulation: the 41K protein Hip (also known as p48) binds specifically to the ATPase domain of mammalian Hsp70, slowing the dissociation of ADP^{64,65}, and may thus stabilize the interaction between Hsp70 and substrate proteins.

Function under stress conditions. Surprisingly little is known about the protective function of Hsp70s under stress conditions, such as the exposure of cells to high temperature. Although a

capacity to solubilize the aggregates of certain heat-denatured proteins has been demonstrated for DnaK *in vitro*⁶⁶, a mechanism in which the chaperone binds to the substrate polypeptide during the stress-induced unfolding, thereby preventing aggregation, may predominate *in vivo*. For example, firefly luciferase expressed in *E. coli* is deactivated when the cells are shifted to 42 °C. Enzyme activity is subsequently regained at 30 °C, but only when functional DnaK, DnaJ and GrpE proteins were present during the inactivation period⁵⁷. In eukaryotic cells, Hsp70 may cooperate with DnaJ homologues and the abundant stress-protein Hsp90 in preventing heat-induced protein denaturation⁶⁷. The function of these components is probably complemented by a class of ATP-independent small heat-shock proteins of ~25K, which may act as a first line of defence against aggregation⁶⁸. Notably, a capacity to solubilize pre-existent protein aggregates *in vivo* has been demonstrated for Hsp104 in yeast, a member of the Hsp100 (Clp) family of stress proteins which help organisms to survive extreme conditions⁶⁹. There is increasing evidence that the function of chaperones in preventing or reversing protein aggregation is linked with their role in presenting misfolded polypeptides to the cellular machinery for proteolytic degradation (reviewed in ref. 70).

Protein translocation across membranes. The Hsp70s in mitochondria and the ER are not only required for the folding of newly translocated proteins, but also for the translocation process itself. The polypeptide chain traverses the mitochondrial membranes through a proteinaceous channel in an extended conformation (Fig. 3). Mitochondrial Hsp70 binds to suitable segments of the translocating chain as it emerges from the inner membrane⁷¹, thus preventing it from sliding backwards⁷² (Fig. 3a). Multiple events of Hsp70 binding and ATP-dependent release would then promote translocation⁷³ in a molecular ratchet-like mechanism⁷⁴. Translocation requires the interaction between Hsp70 and Tim44, an inner mitochondrial membrane protein of the translocation machinery^{73,75,76} which may have a limited homology to DnaJ⁷⁵. Tim44 is thought to present segments of the incoming polypeptide to the ATP-bound form of Hsp70 (Fig. 3a). Recycling of Hsp70 requires the nucleotide-exchange factor Mge1 (ref. 77), the mitochondrial GrpE homologue (Fig. 3b). The post-translational uptake of proteins into the yeast ER seems to rely on a similar mechanism that involves the interaction of the Hsp70 in the ER lumen with the J-domain of Sec63, a membrane component of the translocation machinery⁴⁰.

In addition to its function in protein translocation, Hsp70 fulfills a regulatory role in diverse cellular processes because of its capacity to induce the disassembly of specific protein complexes. Examples include the initiation of DNA synthesis from various phage and plasmid origins of replication in *E. coli* and the uncoating of clathrin-coated endocytotic vesicles in mammalian cells (reviewed in refs 4, 66). In these cases, Hsp70 apparently interacts with proteins that expose chaperone recognition motifs in their folded states.

The chaperonin system

The chaperonins²¹ have an essential function in promoting the ATP-dependent folding of proteins, both under normal growth conditions and under stress. There are two chaperonin subgroups: members of the GroEL (or Hsp60) family, and the chaperonins of the TRiC family (Table 1). (1) The GroEL-type chaperonins in eubacteria, mitochondria and chloroplasts are made of two stacked seven-membered rings of ~60K. They cooperate with a smaller protein cofactor, GroES in *E. coli*, which is a single heptameric ring of ~10K subunits. These proteins are stress-inducible, except for the chloroplast homologue. *E. coli* GroEL has an additional role in the regulation of messenger RNA turnover, which could reflect a link between chaperonins and an ancient RNA world⁷⁸. (2) The TRiC (TCP-1 ring complex) chaperonins in archaeobacteria and the eukaryotic cytosol are 8- or 9-membered double rings containing a family of ~55K subunits with homology to the TCP-1 tail complex protein from mouse. They are apparently independent of a GroES-like cofactor. Only the archaeobacterial members of this family are stress-inducible. Mechanistic studies have concentrated on GroEL and GroES of *E. coli*, which together function as a protein-folding cage. Functional data for TRiC are sparse, but the conservation of the cylindrical structure suggests that it, too, provides a folding compartment. However, in contrast to the GroEL chaperonins, the eukaryotic cytosolic chaperonin appears to have a restricted range of substrates.

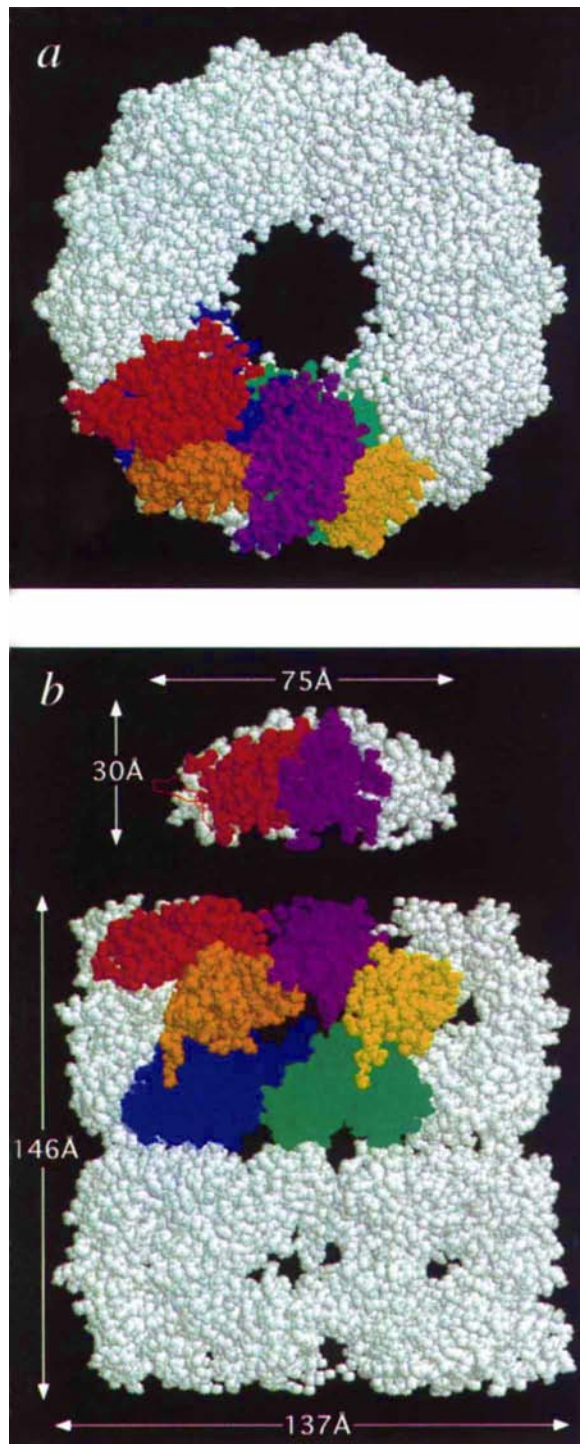
GroEL and GroES structure. GroEL has been widely studied by electron microscopy⁷⁹⁻⁸² and recently by X-ray crystallography^{83,84}. GroEL is a cylindrical complex of ~150 Å height and ~140 Å width that encloses a central cavity ~50 Å wide, the site of polypeptide binding^{79,80,82,85} (Fig. 4a, b). Each 57K subunit is composed of an apical, an intermediate and an equatorial domain. The equatorial domain contains the ATP-binding site⁸⁴ and provides most of the intersubunit interactions (Fig. 4a-c). The small intermediate domain has potential hinge regions at the domain junctions that allow for the movement of the two major domains relative to each other^{82,86}. The apical domains form the flexible opening of the cylinder. Two crystal structures of GroES^{87,88} revealed the dome-shaped structure of the GroES heptamer (Fig. 4b). Each GroES subunit contains a functionally critical loop region that protrudes from the base of the dome and becomes structured upon docking of GroES onto GroEL⁸⁹.

Polypeptide binding. In the absence of GroES, GroEL can typically bind up to two molecules of polypeptide, one in the centre of each of the two rings at the level of the apical domains^{79,80,82,85}. Stable complex formation probably requires that the polypeptide makes contact with several GroEL subunits. Bound protein is in a collapsed, molten globule-like conformation containing unstable secondary structure^{56,90-93}. GroEL recognizes with decreasing affinity a series of equilibrium intermediates of α -lactalbumin, ranging from extended to native-like molten globule species, and may thus assist a protein along most of its folding pathway⁹⁴. The interaction with the substrate seems to be primarily hydrophobic in nature⁹⁴⁻⁹⁶. Indeed, in the crystal structure the apical domains of GroEL expose a number of hydrophobic amino-acid residues towards the cavity, mutation of which abolishes polypeptide binding⁸⁵ (Fig. 4c). The apical domains can undergo an *en bloc* movement relative to the intermediate domains⁸⁶, probably necessary for GroEL to accommodate its large variety of different substrates that range from ~15K to 60K in size⁹⁷⁻⁹⁹.

The chaperonin cycle in folding. Three functional elements contribute to the high efficiency of chaperonin-assisted polypeptide folding: (1) prevention of aggregation by binding unfolded or partially folded polypeptides; (2) release of unfolded polypeptides into an enclosed folding compartment; and (3) rebinding and structural rearrangement of polypeptides that fail to fold. Although the binding of unfolded protein is independent of cofactors, the role of GroEL in providing a folding compartment critically depends on conformational changes that are induced by

GroES. The exit of folded protein into free solution and the rebinding of incompletely folded substrate require the timed cycling of GroES between GroEL-bound and free states, which is controlled by the ATP-hydrolytic activity of GroEL. These aspects will be considered first, followed by a discussion of the distinct steps of the folding cycle.

As revealed by electron microscopy, the asymmetrical binding of GroES to one end-surface of GroEL induces dramatic conformational changes in the interacting GroEL ring^{79,82}. The apical domains in this ring move upwards and outwards relative to the intermediate hinge domains, generating a large enclosed chamber of ~65 Å by 80 Å (ref. 82) (Fig. 4d). Because the hydrophobic binding regions of GroEL for polypeptide overlap with those for GroES⁸⁵, GroES binding leads to the displacement



of the bound protein from its attachment sites on GroEL, forcing it into a more hydrophilic cage which is permissive for folding¹⁰⁰⁻¹⁰³. The association of GroES is sufficient to prevent the protein from diffusing out of the cage, because there is no free passage for polypeptide between GroEL rings^{81,83}.

The binding and unbinding of GroES relies on an interesting network of allosteric interactions, characterized by a positive cooperativity of nucleotide binding and hydrolysis at the level of the individual GroEL rings and a negative cooperativity for ligand binding between rings^{100,104-110}. Initially one ring of GroEL will bind ATP and GroES, thereby directing the formation of asymmetrical GroEL:GroES particles^{79,81} (Fig. 4*d*). ATP hydrolysis in the GroEL subunits that interact with GroES then gives rise to a tight GroEL-7ADP-GroES complex¹⁰⁰. ATP hydrolysis in the opposite ring causes dissociation of the ADP and GroES^{100,107}, hence the requirement for two rings in the functional cycle.

Sufficient information is now available to describe a basic

reaction scheme for GroEL-GroES function, although some details of this model may change in the future (Fig. 4*e*). (1) The acceptor state for unfolded polypeptide is the asymmetrical GroEL-GroES complex which binds substrate in the GroEL ring that is not occupied by GroES^{82,100,111}. (2-3) ATP binding and hydrolysis in this ring (as well as polypeptide binding itself) cause the dissociation of ADP and GroES from the opposite ring^{100,108}. (4) GroES will then rebind, together with ATP, to either GroEL ring, enclosing ~50 per cent of the bound polypeptide in the cavity^{100,102,112}. The enclosed polypeptide initiates folding, whereas binding of substrate and GroES to opposite GroEL rings may result in the release of non-native polypeptide (not shown in the model; see below). (4-5) ATP hydrolysis in the GroES-bound ring tightens the interaction with GroES. (6) ATP binding and hydrolysis in the opposite ring then triggers the opening of the GroEL-GroES cage¹⁰². At this point, the folded polypeptide leaves, whereas incompletely folded polypeptide

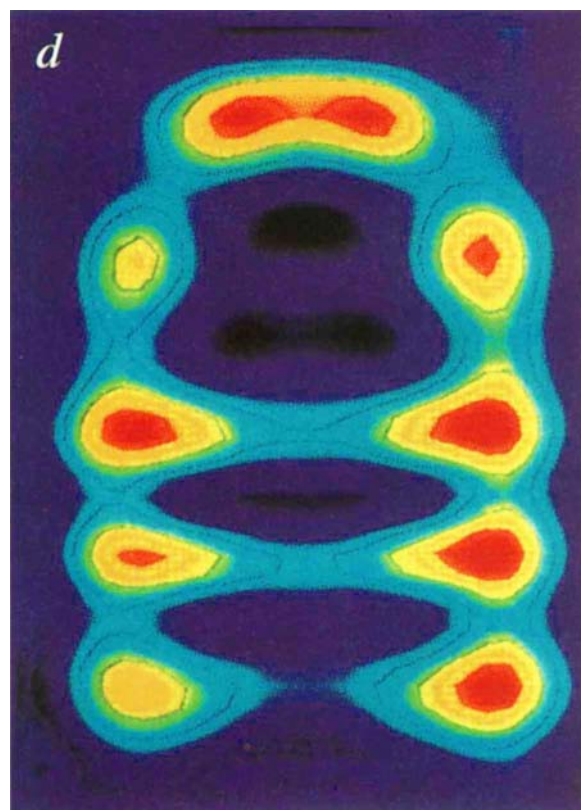
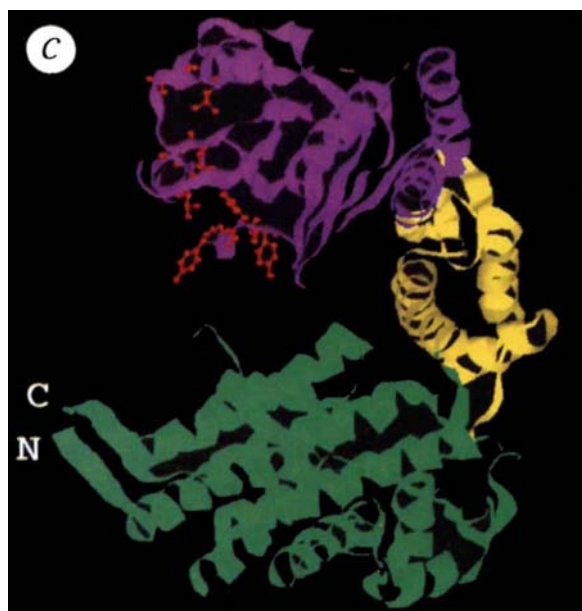
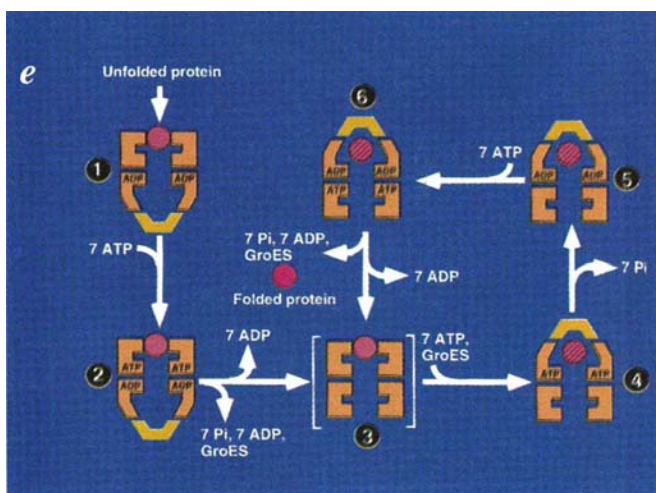


FIG. 4 Structure and function of the chaperonin system. *a, b*, Space-filling representations showing a top and side view, respectively, of the crystal structure of GroEL⁸³. Two adjacent subunits are coloured with the apical domains in red and purple, the intermediate domains in orange and yellow, and the equatorial domains in blue and green, respectively. Free passage between GroEL rings is obstructed by N- and C-terminal residues not resolved in the crystal structure. Side view of the crystal structure of GroES⁸⁷ (shown in *b*). Two adjacent domains and the single mobile loop that is structured in the crystal owing to crystal packing are coloured. The loop regions normally protrude from the base of GroES downwards. *c*, Ribbon diagram of a GroEL subunit showing in red the location of hydrophobic residues (Tyr 199, Ser 201, Tyr 203, Phe 204, Leu 234, 237, 259, Val 263, 264) involved in polypeptide binding on the cavity-facing surface of the apical domain^{85,86}. Apical, intermediate and equatorial domains in purple, yellow and green, respectively. *d*, The asymmetrical GroEL-GroES complex as revealed by cryoelectron microscopy (from ref. 154). Note the upward and outward movement of the apical GroEL domains interacting with GroES. *e*, Model of the GroEL-GroES reaction cycle in folding (from refs 101,102). Nucleotide-free GroEL in step (3) is thought to occur only transiently and is shown for simplicity. In step (4), GroES may either associate with the polypeptide-containing ring of GroEL or with the free ring (not shown). Light and dark pink spheres represent unfolded and folded substrate protein, respectively, and the hatched sphere indicates the presence of a mixture of folded and unfolded substrate in the population of GroEL molecules. Unfolded protein can be retained in steps (6) to (3).



rebinds¹⁰⁰. The energy of rebinding may be used for (partial) unfolding^{5,90,93,102} in preparation for another folding trial. Thus, folding occurs in an iterative series of reactions with first-order kinetics¹⁰², corresponding to multiple ATPase cycles⁹⁰. The fraction of enclosed protein that folds in a single cycle varies, dependent on the folding propensity of the specific protein and the time available between GroES binding and release¹⁰², approximately 20 seconds at 25 °C (refs 107,108).

The release of GroES in steps (2) and (6) of the cycle can occur independently of the binding of a second GroES^{108,109,112} (Fig. 4e). However, the transient formation of a symmetrical chaperonin particle with GroES bound to both ends of GroEL may enhance the efficiency of this reaction. Such complexes can be populated *in vitro* under a specific set of conditions^{111,113–115}. Their functional significance remains to be established.

Folding in the GroEL cage. Folding in the GroEL cavity would be comparable, to a first approximation, to the folding of a single protein molecule in a test tube^{5,116–118}. The basic features of this so-called Anfinsen-cage model^{119,120} have been experimentally established^{100,102,103}. However, placing the folding protein into the confined space of the GroEL cage may have additional physical consequences. Changing the wall of the cylinder from hydrophobic to hydrophilic may drive the burial of hydrophobic residues by the enclosed polypeptide more effectively than in free solution. The rearrangement of not-yet folded protein is a further significant improvement over spontaneous folding *in vitro*^{5,93,102}. It would be equivalent to a proofreading mechanism and may result in catalysis of folding by returning incorrectly folded or kinetically trapped intermediates to a faster-folding track¹²¹ (S. Radford, unpublished).

Strong evidence is now existing in favour of the cage model. Early experiments with purified chaperonin led to the proposal that GroES allows a protein to fold in association with GroEL to a state that is no longer aggregation-prone and is committed to complete folding successfully⁹⁰. Indeed, several monomeric proteins reach their native, enzymatically active conformation in the GroEL cavity^{102,103,122}. Assembly in the case of oligomeric proteins, however, occurs in free solution after the subunits have folded in the chaperonin cage. An alternative view held that GroEL functions by unfolding misfolded polypeptides and ejecting them in a non-native form. Folding would then occur in the bulk solution^{107,112,123}. This was based on the observation that in the presence of ATP and GroES a significant fraction of bound protein can be rapidly released from GroEL in an unfolded state^{107,112,123}. However, it has become clear that this reaction is not productive for folding^{102,103} and rather represents a premature loss of protein that may occur when GroES and substrate protein are bound to opposite GroEL rings (see above) or when GroES unbinds. Although the magnitude of this effect in the intact *E. coli* cell is unknown, the possibility of releasing non-native protein from GroEL is probably relevant for providing an exit route for aberrant polypeptides that must be degraded¹⁰² or in allowing the transient interaction with polypeptides that are too large to be enclosed by GroES.

Chaperonin specificity: GroEL versus TRiC. A functional comparison of GroEL with TRiC, the chaperonin of the eukaryotic cytosol (also known as CCT¹²⁴), has shown that the chaperonin mechanism is more specific than previously assumed. TRiC is required *in vivo* for the folding of a subset of polypeptides, including actin and tubulin^{125,126}, and folds firefly luciferase *in vitro*^{35,127}. Interestingly, both unfolded actin and luciferase form a complex with GroEL, but will not fold in the presence of ATP and GroES^{127,128}. Only TRiC has the capacity of generating a set of folding intermediates that allow these proteins to proceed to the native state. TRiC differs from GroEL in that it contains up to eight distinct subunits¹²⁴. Although the general architecture and the overall domain structure of the subunits is conserved between both chaperonins^{125,127,129–132}, little sequence conservation is detected between the apical polypeptide-binding domains^{131,132}. This is also the region where the individual TRiC subunits deviate

most from one another. Thus, TRiC may have coevolved to stabilize the productive folding intermediates of certain eukaryotic proteins. As folding apparently occurs in association with TRiC^{128,133}, the absence of a GroES raises the question of how the substrate is prevented from leaving the chaperonin cavity prematurely. Perhaps release from TRiC involves a stepwise dissociation of the polypeptide by domains or subdomains.

Cellular protein folding pathways

An important question concerns the extent to which molecular chaperones are actually involved in *de novo* protein folding and how different chaperones cooperate *in vivo*. A small number of recent studies, carried out in whole cells or in translation extracts, provide some information about the folding pathways of specific newly synthesized polypeptides and about the mechanisms that ensure their orderly interaction with defined sets of chaperones. A brief description of protein folding in the cytosol and in mitochondria may serve to illustrate this. A discussion of protein folding in the ER is beyond the scope of this article (reviewed in refs 4, 134). **Bacterial cytosol and mitochondrial matrix: homologous folding compartments.** Both DnaK/DnaJ/GrpE and GroEL/GroES must be expressed in the *E. coli* cytosol to prevent the aggregation of a large fraction of newly synthesized polypeptides, even under non-stress conditions¹³⁵. These components constitute a minimal set of chaperones which are typically conserved, even in small bacterial genomes¹³⁶. Both systems interact with similar subsets of proteins, consistent with their proposed functional cooperation²⁶. DnaJ and DnaK, but not GroEL, have been found in association with nascent chains and ribosomes *in vivo* and *in vitro*^{33,34,137,138}. Efficient folding of rhodanese upon cell-free translation can be dependent on all five chaperone components, whereby DnaJ and DnaK must bind the polypeptide before its release from the ribosome to allow efficient transfer to GroEL¹³⁹ (J. Hendrick and F.U.H., unpublished) (Fig. 5). The interaction with DnaJ/DnaK may be preceded by the binding of trigger factor (TF) to the nascent chain^{140,141}, a 48K protein with ribosome affinity that was initially described as a putative chaperone for secretory proteins¹⁴². TF has prolyl isomerase activity *in vitro* and may act as both chaperone and folding catalyst *in vivo*¹⁴³. An interaction between TF and GroEL has also been reported¹⁴⁴.

It is still unclear what fraction of cytosolic polypeptides have to interact with chaperones during translation, and the protective function of the ribosome itself needs to be fully investigated. The concentration of potential nascent chain-binding proteins is at least equivalent to that of ribosomes, but the basal level of GroEL is much lower (~4 μM, compared to ~35 μM ribosomes)³⁰. Thus, the post-translational interaction with GroEL may be restricted to those proteins, which are highly aggregation-sensitive. Consistent with this, under normal growth conditions, the folding of ~30% of all cytosolic protein species was affected in a GroEL mutant strain⁹⁹. Understanding the common structural properties of the *in vivo* substrates of GroEL and the mechanisms by which they are delivered to the chaperonin will be of considerable interest. Other questions concern the directionality of polypeptide transfer between different chaperones during folding and assembly. For example, the subunits of oligomeric proteins may engage in a renewed interaction with DnaK/DnaJ after their folding on GroEL. A retrograde transfer from the chaperonin to the Hsp70 system may also be involved in the targeting of polypeptides for degradation⁷⁰ (Fig. 5). Finally, how do large cytosolic proteins fold, such as β-galactosidase, which do not fit into the central cavity of GroEL?

Some of these questions have been addressed for the folding of proteins in the mitochondrial matrix, the evolutionary equivalent of the bacterial cytosol. Upon import into the organelles, all proteins interact first with mitochondrial Hsp70 (Fig. 3), a reaction that may be analogous to the binding of cytosolic Hsp70 to nascent chains³². In a process dependent on mitochondrial DnaJ and GrpE^{77,145}, aggregation-sensitive proteins must subsequently be transferred to the chaperonin Hsp60 for folding to the native

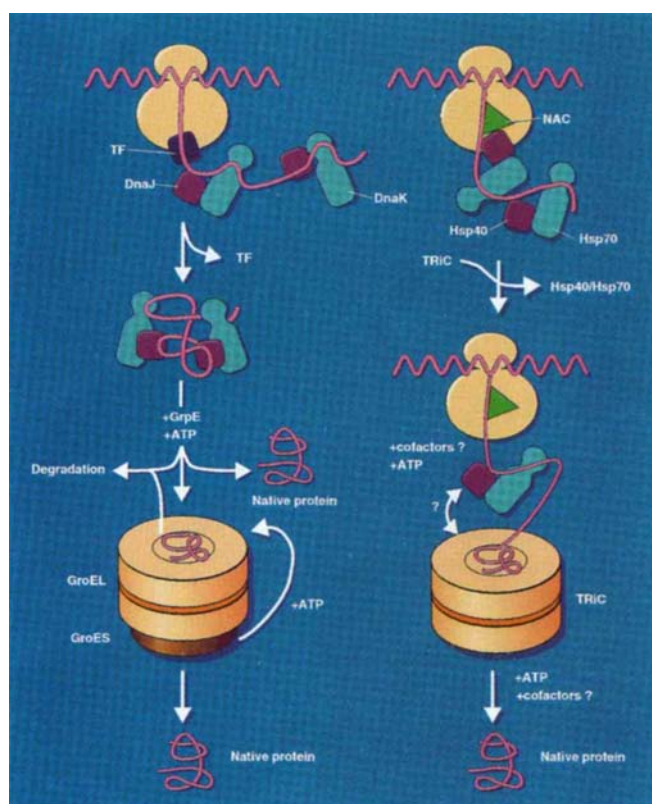


FIG. 5 Possible chaperone pathways for the folding of cytosolic proteins in bacteria (left) and in mammalian cells (right). Although alternative routes for folding are likely to exist, these pathways are thought to be involved in the folding of a subset of aggregation-sensitive polypeptides (see text). The possibility of folding without participation of chaperones in a scheme as shown in Fig. 1b and c is not excluded. Left, DnaJ and DnaK may also interact at a late stage of translation or post-translationally (only cotranslational interaction is shown). Aberrant proteins that are unable to fold may be presented to the proteolytic degradation machinery either from DnaK/DnaJ or from GroEL (shown as release from the GroEL ring opposite GroES). Folded monomeric proteins or the folded subunits of oligomeric protein complexes are released from underneath GroES. TF, trigger factor. Right, proposed pathway for the folding of actin and firefly luciferase, based on studies in cell-free extracts^{35,133}. The action of Hsp70/Hsp40 and TRiC may be functionally coupled. A polypeptide consisting of more than one domain may initiate folding cotranslationally.

state⁵. The directionality of this pathway, established in reconstitution experiments with mitochondrial rhodanese *in vitro*²⁶ (Fig. 2d), has been confirmed for rhodanese in intact organelles¹⁴⁶. In contrast, proteins that refold spontaneously *in vitro* with high efficiency, such as the normally cytosolic protein dihydrofolate reductase (DHFR) from mouse, can fold independently of Hsp60, at least in mitochondria of the yeast *Saccharomyces cerevisiae* at 25 °C (refs 146, 147). Folding of these proteins may however become Hsp60-dependent at higher temperature⁹⁸. Interestingly, in the mitochondria of another fungus, *Neurospora crassa*, DHFR interacts efficiently with Hsp60 during folding even at 25 °C (refs 25, 148), and this interaction is prolonged in organelles defective in a prolyl isomerase that accelerates DHFR folding¹⁴⁸. Thus, not only the intrinsic folding propensity of a given protein and the protein environment, but also the binding properties of the chaperones it encounters may determine the extent to which the chaperone machinery participates in folding *in vivo*.

Folding in the eukaryotic cytosol. A number of components have been found to associate with nascent chains in the mammalian cytosol and are thought to be involved in their folding, including the nascent-chain-associated complex (NAC), the signal-recognition

particle (SRP), Hsp70/Hsc70, the DnaJ homologue Hsp40, and the chaperonin TRiC. Although most chains seem to interact cotranslationally with Hsp70/DnaJ^{15,32,35}, the chaperonin TRiC apparently mediates only the folding of a limited set of proteins, as suggested by its low abundance in many cell types⁵. Thus, how the majority of cytosolic proteins in eukaryotes fold at the chaperonin step remains to be established. There is no evidence that other abundant chaperones, such as Hsp90, have a general role in *de novo* folding⁶⁸, and it is possible that only a small fraction of eukaryotic proteins need a chaperonin.

A cooperative chaperone pathway involving NAC, Hsc70/Hsp40 and TRiC has been proposed for the folding of firefly luciferase and actin based on experiments in a mammalian translation extract^{35,133,149}. This pathway may be important *in vivo* for substrates such as actin and tubulin, whose folding is TRiC dependent. A model proposed to describe the sequence of steps in this reaction is shown in Fig. 5. (1) The first cytosolic factor thought to be encountered by many nascent chains is NAC, a heterodimer of 33K and 21K subunits with an affinity for ribosomes^{149,150}. NAC shields the ~33 residues of a nascent polypeptide adjacent to the peptidyl transferase site and prevents the association of translating ribosomes with the ER membrane, except when a secretory protein destined for translocation into the ER is synthesized that is recognized by SRP. (2) NAC binding may signal that a nascent polypeptide is about to emerge from the ribosome and could be involved in recruiting the chaperone components required for further protection of the growing polypeptide. In any case, Hsc70 and Hsp40 can associate with nascent chains that expose less than 50 amino acid residues beyond the NAC binding site³⁵. Hsp40 activates the Hsc70 ATPase^{64,151} and has a catalytic function in loading Hsc70 onto the nascent chain³⁵ (Fig. 2c). (3) Unlike GroEL, TRiC has the capacity to bind to nascent polypeptides^{35,133} (H. Sternlicht, unpublished), at least in an *in vitro* translation extract. TRiC can be recruited to the elongating chain when ~150–200 amino-acid residues have emerged from the ribosome, consistent with a role of TRiC in mediating the cotranslational folding of certain polypeptide domains³⁵ (Fig. 5). However, completion of folding is only observed after the release of the full-length polypeptide from the ribosome^{35,37,133} (Fig. 1). At this point, the polypeptide–TRiC complex dissociates in an ATP-dependent reaction, resulting in the rapid formation of the native protein. Future research will have to reveal the mechanisms by which the different chaperones are specifically recruited to the translating polypeptide and the degree of their functional coupling in this pathway.

Perspectives

The principles of protein refolding established *in vitro* also govern the *de novo* folding of proteins *in vivo*. This has become clear in particular from recent studies of the chaperonin mechanism. However, the requirement of a highly coordinated protein machinery to assist a polypeptide along most of its cellular folding pathway provides a fascinating additional layer of complexity that warrants detailed analysis. It must be emphasized that proteins and the cellular folding machinery have coevolved. Thus, using chaperone components as tools, much may be learned about the properties of critical folding intermediates and the rules of the folding process.

Another focus of future research is likely to concern the possible role of molecular chaperones in various pathological states. The basic defect in a number of human genetic diseases is known to affect the folding and intracellular trafficking of key proteins¹⁵², cystic fibrosis being perhaps the most prominent example. Molecular chaperones may also be involved in the pathogenesis of other protein-folding defects, including certain prion diseases and perhaps amyloidosis. □

F. Ulrich Hartl is in the Howard Hughes Medical Institute and Cellular Biochemistry and Biophysics Program, Memorial Sloan-Kettering Cancer Center, 1275 York Avenue, New York, New York 10021, USA

1. Anfinsen, C. B. *Science* **181**, 223–230 (1973).
2. Ellis, R. J. *Nature* **328**, 378–379 (1987).
3. Rothman, J. E. *Cell* **59**, 591–601 (1989).
4. Gething, M.-J. & Sambrook, J. *Nature* **355**, 33–45 (1992).
5. Hendrick, J. P. & Hartl, F. U. *Rev. Biochem. Sci.* **62**, 349–384 (1993).
6. Freedman, R. B., Hirst, T. R. & Tuit, M. F. *Trends biochem. Sci.* **19**, 331–336 (1994).
7. Schmid, F. X. A. *Rev. Biophys. biomol. Struct.* **22**, 123–143 (1993).
8. Laskey, R. A., Honda, B. M. & Finch, J. T. *Nature* **275**, 416–420 (1978).
9. Hightower, L. E. *J. Cell Physiol.* **102**, 407–427 (1980).
10. Pelham, H. R. B. *Cell* **46**, 959–961 (1986).
11. Haas, I. G. & Wabl, M. *Nature* **306**, 387–389 (1983).
12. Deshaies, R. J., Koch, B. D., Werner-Washburne, M., Craig, E. A. & Schekman, R. *Nature* **332**, 800–805 (1988).
13. Chirico, W. J., Waters, M. G. & Blobel, G. *Nature* **332**, 805–810 (1988).
14. Zimmermann, R., Sagstetter, M., Lewis, M. J. & Pelham, H. R. B. *EMBO J.* **7**, 2875–2880 (1988).
15. Beckmann, R. P., Mizzen, L. A. & Welch, W. J. *Science* **248**, 850–854 (1990).
16. Flynn, G. C., Pohl, J., Flocco, M. T. & Rothman, J. E. *Nature* **353**, 726–730 (1991).
17. Georgopoulos, C., Hendrix, R. W., Casjens, S. R. & Kaiser, A. D. *J. molec. Biol.* **76**, 45–60 (1973).
18. Coppo, A., Manzi, A., Pulitzer, J. F. & Takahashi, H. J. *molec. Biol.* **76**, 61–87 (1973).
19. Barraclough, R. & Ellis, R. J. *Biochim. biophys. Acta* **608**, 18–31 (1980).
20. Cannon, S., Wang, P. & Roy, H. J. *Cell Biol.* **103**, 1327–1335 (1986).
21. Hemmingsen, S. M. *et al. Nature* **333**, 330–334 (1988).
22. McMullin, T. W. & Hallberg, R. L. *Molec. cell. Biol.* **8**, 371–380 (1988).
23. Goloubinoff, P., Christeller, J. T., Gatenby, A. A. & Lorimer, G. H. *Nature* **342**, 884–889 (1989).
24. Cheng, M. Y. *et al. Nature* **337**, 620–625 (1989).
25. Ostermann, J., Horwich, A. L., Neupert, W. & Hartl, F. U. *Nature* **341**, 125–130 (1989).
26. Langer, T. *et al. Nature* **356**, 683–689 (1992).
27. Creighton, T. E. *Proteins: Structures and Molecular Properties* (Freeman, New York, 1993).
28. Dobson, C. M., Radford, S. E. & Evans, P. A. *Trends biochem. Sci.* **19**, 31–37 (1994).
29. Mitrali, A., Fane, B., Haase-Pettingill, C., Sturtevant, J. & King, J. *Science* **253**, 54–58 (1991).
30. Ellis, R. J. & Hartl, F. U. *FASEB J.* **10**, 20–26 (1996).
31. Zimmermann, S. B. & Trach, S. O. *J. molec. Biol.* **222**, 599–620 (1991).
32. Nelson, R. J., Ziegelhoffer, T., Nicolet, C., Werner-Washburne, M. & Craig, E. A. *Cell* **71**, 97–105 (1993).
33. Kudlicki, W. *et al. J. biol. Chem.* **270**, 10650–10657 (1995).
34. Hendrick, J. P., Langer, T., Davis, T. A., Hartl, F. U. & Wiedmann, M. *Proc. natn. Acad. Sci. U.S.A.* **90**, 10216–10220 (1993).
35. Frydman, J., Nimmesgern, E., Ohtsuka, K. & Hartl, F. U. *Nature* **370**, 111–117 (1994).
36. Chattopadhyay, S., Das, B., Bera, A. K., Dasgupta, D. & Dasgupta, C. *Biochem. J.* **300**, 717–721 (1994).
37. Kolb, V. A., Makeyev, E. V. & Spirin, A. S. *EMBO J.* **13**, 3631–3637 (1994).
38. Fedorov, A. N. & Baldwin, T. O. *Proc. natn. Acad. Sci. U.S.A.* **92**, 1227–1231 (1995).
39. Chen, W., Helenius, J., Braakman, I. & Helenius, A. *Proc. natn. Acad. Sci. U.S.A.* **92**, 6229–6233 (1995).
40. Sanders, S. L., Whitfield, K. M., Vogel, J. P., Rose, M. D. & Schekman, R. W. *Cell* **69**, 353–365 (1992).
41. Ungewickell, E. *et al. Nature* **378**, 632–635 (1995).
42. Schlenstedt, G., Harris, S., Risse, B., Lill, R. & Silver, P. A. *J. Cell Biol.* **129**, 979–988 (1995).
43. Flaherty, K. M., Deluca-Flaherty, C. & McKay, D. B. *Nature* **346**, 623–628 (1990).
44. Liberek, K., Skowry, D., Zyllicz, M., Johnson, C. & Georgopoulos, C. *J. biol. Chem.* **266**, 14491–14496 (1991).
45. Buchberger, A., Schröder, H., Buttner, M., Valencia, A. & Bukau, B. *Nature struct. Biol.* **1**, 95–101 (1994).
46. Morshauer, R. C., Wang, H., Flynn, G. C. & Zwieterweg, E. R. P. *Biochemistry* **34**, 6261–6266 (1995).
47. Wang, T.-F., Chang, J.-H. & Wang, C. *J. biol. Chem.* **268**, 26049–26051 (1993).
48. Szyperki, T., Pellicchia, M., Wall, D., Georgopoulos, C. & Wuthrich, K. *Proc. natn. Acad. Sci. U.S.A.* **91**, 11343–11347 (1994).
49. Hill, R. B., Flanagan, J. M. & Prestegard, J. H. *Biochemistry* **34**, 5587–5596 (1995).
50. Wall, D., Zyllicz, M. & Georgopoulos, C. *J. biol. Chem.* **269**, 5446–5451 (1994).
51. Szabo, A., Korszun, R., Hartl, F. U. & Flanagan, J. *EMBO J.* **15**, 408–417 (1996).
52. Blond-Elguindi, S. *et al. Cell* **75**, 717–728 (1993).
53. Gragerov, A. & Gottesman, M. E. *J. molec. Biol.* **241**, 133–135 (1994).
54. Takenaka, I. M., Leung, S. M., McAndrew, S. J., Brown, J. P. & Hightower, L. E. *J. biol. Chem.* **270**, 19839–19844 (1995).
55. Palleros, D. R., Welch, W. J. & Fink, A. L. *Proc. natn. Acad. Sci. U.S.A.* **88**, 5719–5723 (1991).
56. Landry, S. J., Jordan, R., McMacken, R. & Gierasch, L. M. *Nature* **355**, 455–457 (1992).
57. Schröder, H., Langer, T., Hartl, F. U. & Bukau, B. *EMBO J.* **12**, 4137–4144 (1993).
58. Kzabo, A. *et al. Proc. natn. Acad. Sci. U.S.A.* **91**, 10345–10349 (1994).
59. Schmid, D., Baici, A., Gehring, H. & Christen, P. *Science* **263**, 971–973 (1994).
60. Liberek, K., Marszałek, J., Ang, D., Georgopoulos, D. & Zyllicz, M. *Proc. natn. Acad. Sci. U.S.A.* **88**, 2874–2878 (1991).
61. Palleros, D. R., Reid, K. L., Shi, L., Welch, W. J. & Fink, A. L. *Nature* **365**, 664–666 (1993).
62. Palleros, D. R., Shi, L., Reid, K. L. & Fink, A. L. *J. biol. Chem.* **269**, 13107–13114 (1994).
63. Ziegelhoffer, T., Lopezbuesca, P. & Craig, E. A. *J. biol. Chem.* **270**, 10412–10419 (1995).
64. Höfheid, J., Minami, Y. & Hartl, F. U. *Cell* **83**, 589–598 (1995).
65. Prapanich, V., Chen, S., Nair, S., Rimmerman, R. A. & Smith, D. F. *Molec. Endocr.* **11**, 420–428 (1996).
66. Georgopoulos, C., Liberek, K., Zyllicz, M. & Ang, D. *Heat-shock Proteins in Biology and Medicine* 209–249 (Cold Spring Harbor Laboratory Press, Cold Spring Harbor, New York, 1994).
67. Freeman, B. C. & Morimoto, R. I. *EMBO J.* **12**, 2969–2979 (1996).
68. Jakob, U. & Buchner, J. *Trends biochem. Sci.* **19**, 205–211 (1994).
69. Parsell, D. N., Kowal, A. S., Singer, M. A. & Lindquist, S. *Nature* **372**, 475–478 (1994).
70. Hayes, S. A. & Dice, F. J. *Cell Biol.* **132**, 255–258 (1996).
71. Kang, P. J. *et al. Nature* **348**, 137–143 (1990).
72. Ungermann, C., Neupert, W. & Cyr, D. M. *Science* **266**, 1250–1253 (1994).
73. Schneider, H. C. *et al. Nature* **371**, 768–774 (1994).
74. Simon, S. M., Peskin, C. S. & Oster, G. F. *Proc. natn. Acad. Sci. U.S.A.* **89**, 3770–3774 (1992).
75. Rassow, J. *et al. J. Cell Biol.* **127**, 1547–1556 (1994).
76. Kronidou, N. G. *et al. Proc. natn. Acad. Sci. U.S.A.* **91**, 12818–12822 (1994).
77. Westermann, B., Pripbuus, C., Neupert, W. & Schwarz, E. *EMBO J.* **14**, 3452–3460 (1995).
78. Georgellis, D., Sohlberg, B., Hartl, F. U. & von Gabain, A. *Molec. Microbiol.* **16**, 1259–1268 (1995).
79. Langer, T., Pfeifer, G., Martin, J., Baumeister, W. & Hartl, F. U. *EMBO J.* **11**, 4757–4765 (1992).
80. Braig, K., Simon, M., Furuya, F., Hainfeld, J. F. & Horwich, A. L. *Proc. natn. Acad. Sci. U.S.A.* **90**, 3978–3982 (1993).
81. Saibil, H. R. *et al. Curr. Biol.* **3**, 265–273 (1993).
82. Chen, S. *et al. Nature* **371**, 261–264 (1994).
83. Braig, K. *et al. Nature* **371**, 578–586 (1994).
84. Boisvert, D. C., Wang, J., Otwinowski, Z., Horwich, A. L. & Sigler, P. B. *Nature struct. Biol.* **3**, 170–177 (1996).
85. Fenton, W. A., Kashi, Y., Furtak, K. & Horwich, A. L. *Nature* **371**, 614–619 (1994).
86. Braig, K., Adams, P. D. & Brünner, A. T. *Nature struct. Biol.* **2**, 1083–1084 (1995).
87. Hunt, J. F., Weaver, A. J., Landry, S. J., Gierasch, L. & Deisenhofer, J. *Nature* **379**, 37–42 (1996).
88. Mande, S. C., Mehra, V., Bloom, B. & Hol, W. G. J. *Science* **271**, 203–207 (1996).
89. Landry, S. J., Zeilstra, R. J., Fayet, O., Georgopoulos, C. & Gierasch, L. M. *Nature* **364**, 255–258 (1993).
90. Martin, J. *et al. Nature* **352**, 36–42 (1991).
91. Robinson, C. V. *et al. Nature* **372**, 646–651 (1994).
92. Zahn, R., Spitzfaden, C., Ottiger, M., Wuthrich, K. & Pluckthun, A. *Nature* **368**, 261–265 (1994).
93. Zahn, R., Perret, S., Stenberg, G. & Fersht, A. R. *Science* **271**, 642–645 (1996).
94. Hayer-Hartl, M. K., Ewbank, J. J., Creighton, T. E. & Hartl, F. U. *EMBO J.* **13**, 3192–3202 (1994).
95. Lin, Z., Schwarz, F. P. & Eisenstein, E. J. *biol. Chem.* **270**, 1011–1014 (1995).
96. Hlodan, R., Tempst, P. & Hartl, F. U. *Nature struct. Biol.* **2**, 587–595 (1995).
97. Viitanen, P. V., Gatenby, A. A. & Lorimer, G. H. *Protein Sci.* **1**, 363–369 (1992).
98. Martin, J., Horwich, A. L. & Hartl, F. U. *Science* **258**, 995–998 (1992).
99. Horwich, A. L., Low, K. B., Fenton, W. A., Hirshfield, I. N. & Furtak, K. *Cell* **74**, 909–917 (1993).
100. Martin, J., Mayhew, M., Langer, T. & Hartl, F. U. *Nature* **366**, 228–233 (1993).
101. Hartl, F. U. *Nature* **371**, 557–559 (1994).
102. Mayhew, M. *et al. Nature* **379**, 420–426 (1996).
103. Weissman, J. S., Rye, H. S., Fenton, W. A., Beechem, J. M. & Horwich, A. L. *Cell* **84**, 481–490 (1996).
104. Gray, T. E. & Fersht, A. R. *FEBS Lett.* **292**, 254–258 (1991).
105. Todd, M. J., Viitanen, P. V. & Lorimer, G. H. *Biochemistry* **32**, 8560–8567 (1993).
106. Jackson, G. S. *et al. Biochemistry* **32**, 2554–2563 (1993).
107. Todd, M. J., Viitanen, P. V. & Lorimer, G. H. *Science* **265**, 659–666 (1994).
108. Hayer-Hartl, M. K., Martin, J. & Hartl, F. U. *Science* **269**, 836–841 (1995).
109. Burston, S. G., Ranson, N. A. & Clarke, A. R. *J. molec. Biol.* **249**, 138–152 (1995).
110. Yifrach, O. & Horowitz, A. *Biochemistry* **34**, 5303–5308 (1995).
111. Engel, A. *et al. Science* **269**, 832–836 (1995).
112. Weissman, J. S. *et al. Cell* **83**, 577–587 (1995).
113. Llorca, O., Marco, S., Carrascosa, J. L. & Valpuesta, J. M. *FEBS Lett.* **345**, 181–186 (1994).
114. Schmidt, M. *et al. Science* **265**, 656–659 (1994).
115. Azem, A., Kessel, M. & Goloubinoff, P. *Science* **265**, 653–656 (1994).
116. Nilsson, B. & Anderson, S. A. *Rev. Microbiol.* **45**, 607–635 (1991).
117. Creighton, T. E. *Nature* **352**, 17–18 (1991).
118. Agard, D. A. *Science* **260**, 1903–1904 (1993).
119. Ellis, R. J. *Curr. Biol.* **4**, 633–635 (1994).
120. Ellis, R. J. *Fold. Des.* **1**, R9–R15 (1996).
121. Ranson, N. A., Dunster, N. J., Burston, S. G. & Clarke, A. R. *J. molec. Biol.* **250**, 581–586 (1995).
122. Gray, T. E. & Fersht, A. R. *J. molec. Biol.* **232**, 1197–1207 (1993).
123. Weissman, J. S., Kashi, Y., Fenton, W. A. & Horwich, A. L. *Cell* **78**, 693–702 (1994).
124. Kubota, H., Hynes, G., Carne, A., Ashworth, A. & Willison, K. *Curr. Biol.* **4**, 89–99 (1994).
125. Gao, Y., Thomas, J. O., Chow, R. L., Lee, G. H. & Cowan, N. J. *Cell* **69**, 1043–1050 (1992).
126. Yaffe, M. B. *et al. Nature* **358**, 245–248 (1992).
127. Frydman, J. *et al. EMBO J.* **11**, 4767–4778 (1992).
128. Tian, G. L., Vainberg, I. E., Tap, W. D., Lewis, S. A. & Cowan, N. J. *Nature* **375**, 250–253 (1995).
129. Trent, J. D., Nimmesgern, E., Wall, J. S., Hartl, F. U. & Horwich, A. L. *Nature* **354**, 490–493 (1991).
130. Phipps, B. M. *et al. Nature* **361**, 475–477 (1993).
131. Kim, S., Willison, K. R. & Horwich, A. L. *Trends biochem. Sci.* **19**, 543–548 (1994).
132. Waldmann, T., Lupas, A., Kellermann, J., Peters, J. & Baumeister, W. *Biol. Chem. Hoppe Seyler* **376**, 119–126 (1995).
133. Frydman, J. & Hartl, F. U. *Science* **272**, 1497–1502 (1996).
134. Bergeron, J. J., Brenner, M. B., Thomas, D. Y. & Williams, D. B. *Trends biochem. Sci.* **19**, 124–128 (1994).
135. Gragerov, A. *et al. Proc. natn. Acad. Sci. U.S.A.* **89**, 10341–10344 (1992).
136. Koonin, E. V., Mushegian, A. R. & Rudd, K. E. *Curr. Biol.* **6**, 404–416 (1996).
137. Gaitanaris, G. A., Vysokanov, A., Hung, S.-Z., Gottesman, M. & Gragerov, A. *Molec. Microbiol.* **14**, 861–869 (1994).
138. Vysokanov, A. V. *FEBS Lett.* **375**, 211–214 (1995).
139. Kudlicki, W., Odom, O. W., Kramer, G. & Hardesty, B. J. *molec. Biol.* **244**, 319–331 (1994).
140. Valent, Q. A. *et al. EMBO J.* **14**, 5494–5505 (1995).
141. Hesterkamp, T., Hauser, S., Lütcke, H. & Bukau, B. *Proc. natn. Acad. Sci. U.S.A.* **93**, 4437–4441 (1996).
142. Crooke, E. & Wickner, W. *Proc. natn. Acad. Sci. U.S.A.* **84**, 5216–5220 (1987).
143. Stoller, G. *et al. EMBO J.* **14**, 4939–4948 (1995).
144. Kandrö, O., Sherman, M., Rhode, M. & Goldberg, A. L. *EMBO J.* **14**, 6021–6027 (1995).
145. Rowley, N. *et al. Cell* **77**, 249–259 (1994).
146. Rospert, S. *et al. EMBO J.* **15**, 764–774 (1996).
147. Höfheid, J. & Hartl, F. U. *J. Cell Biol.* **126**, 305–315 (1994).
148. Rassow, J. *et al. Molec. cell. Biol.* **15**, 2654–2662 (1995).
149. Wiedmann, B., Sakai, H., Davis, T. A. & Wiedmann, M. *Nature* **370**, 434–440 (1994).
150. Wang, S., Sakai, H. & Wiedmann, M. *J. Cell Biol.* **130**, 519–527 (1995).
151. Freeman, B. C., Myers, M. P., Schumacher, R. & Morimoto, R. I. *EMBO J.* **14**, 2281–2292 (1995).
152. Amara, J. F., Cheng, S. H. & Smith, A. E. *Trends Cell Biol.* **2**, 145–149 (1992).
153. Dobson, C. M. *Curr. Biol.* **2**, 343–345 (1992).
154. Saibil, H. R. *Nature struct. Biol.* **1**, 838–842 (1994).

ACKNOWLEDGEMENTS. Limited space has meant that it has not been possible to cite all relevant references. I thank F. Ilchert and M. Mayhew for preparing the figures; B. Netzer, J. Hendrick, J. Frydman and M. Hartl for critically reading the manuscript; T. Creighton, J. Ellis, C. Dobson, J. Hunt and S. Radford for discussion; H. Deisenhofer and J. Hunt for the crystallographic coordinates of GroES; and R. Morimoto, W. Neupert, H. Sternlicht and S. Radford for sharing unpublished data.

A 3.5-Gyr-old galaxy at redshift 1.55

James Dunlop*, John Peacock†, Hyron Spinrad‡, Arjun Dey§, Raul Jimenez†, Daniel Stern‡ & Rogier Windhorst||

* Institute for Astronomy, Department of Physics and Astronomy, The University of Edinburgh, Edinburgh EH9 3HJ, UK

† Royal Observatory, Edinburgh EH9 3HJ, UK

‡ Astronomy Department, University of California, Berkeley, California 94720, USA

§ NOAO/KPNO, 950 North Cherry Avenue, Tucson, Arizona 85726, USA

|| Department of Physics and Astronomy, Arizona State University, Tempe, Arizona 85287-1504, USA

ONE of the most direct methods of constraining the epoch at which the first galaxies formed—and thereby to constrain the age of the Universe—is to identify and date the oldest galaxies at high redshift. But most distant galaxies have been identified on the basis of their abnormal brightness in some spectral region^{1–4}; such selection criteria are biased towards objects with pronounced nuclear activity or young star-forming systems, in which the spectral signature of older stellar populations will be concealed. Here we report the discovery of a weak and extremely red radio galaxy (53W091) at $z = 1.55$, and present spectroscopic evidence that its red colour results from a population of old stars. Comparing our spectral data with models of the evolution of stellar populations, we estimate that we are observing this galaxy at least 3.5 Gyr after star-formation activity ceased. This implies an extremely high formation redshift ($z > 4$) for 53W091 and, by inference, other elliptical galaxies. Moreover, the age of 53W091 is greater than the predicted age of the Universe at $z = 1.55$, under the assumption of a standard Einstein–de Sitter cosmology (for any Hubble constant greater than $50 \text{ km s}^{-1} \text{ Mpc}^{-1}$), indicating that this cosmological model can be formally excluded.

It is relatively easy for a short-lived burst of star-formation⁵, or re-processed light from an active nucleus^{6,7}, to dominate the appearance of a galaxy in the rest-frame optical–ultraviolet, concealing the presence of an older underlying stellar population⁸. It is thus the reddest objects at a given redshift which are of greatest importance for constraining the first epoch of galaxy formation^{9–11}, and the correlation between the ultraviolet and radio properties of powerful radio galaxies^{12,13} indicates that radio-based searches for passive elliptical galaxies at $z > 1$ should be confined to millijansky flux-density levels. We are therefore investigating the properties of radio galaxies with flux densities at 1.4 GHz $S_{1.4\text{GHz}} > 1 \text{ mJy}$ selected from the Leiden–Berkeley deep survey (LBDS) and its extensions^{14–17}. We now possess optical–infrared photometry down to $V \approx 26 \text{ mag}$ and $K \approx 20 \text{ mag}$ for a complete sample of 77 galaxies, enabling us to estimate redshifts both from spectral fitting¹³ and from a modified version of the infrared Hubble diagram¹⁸. From this sample we have isolated a subset of 10 extremely red ($R - K > 5$) potentially high-redshift (estimated redshift $z_{\text{est}} > 1.5$) objects for intensive spectroscopic study.

The galaxy 53W091, one of the reddest in the sample ($R = 24.6 \pm 0.20 \text{ mag}$ and $K = 18.75 \pm 0.05 \text{ mag}$ within a 4-arcsec aperture) was observed for 1.5 hours on 25 July 1995 with the Low Resolution Imaging Spectrograph (LRIS) on the 10-m W. M. Keck Telescope on Mauna Kea, Hawaii. This observation yielded the detection of a faint and very red continuum, but no emission-line redshift. Therefore, in an attempt to constrain the galaxy redshift, we obtained deep J- and H-band images with IRCAM3 on the United Kingdom Infrared Telescope in August 1995. The ease with which the galaxy was detected at J and H

($J = 20.5 \pm 0.1 \text{ mag}$, $H = 19.5 \pm 0.1 \text{ mag}$; Fig. 1a) revealed that, if its red $R - K$ colour was in part due to a redshifted 4,000 Å break (the most prominent spectral feature displayed by an old stellar population at optical wavelengths), this feature must lie at observed wavelength $\lambda_{\text{obs}} < 1.2 \mu\text{m}$, constraining the redshift to $z < 2$.

Finally, on 31 August and 1 September 1995 we re-observed 53W091 for a total of 4 hours, again with LRIS on the 10-m Keck Telescope (Fig. 1b). The spectrum produced by co-adding all 5.5 hours of integration is shown in Fig. 1c, plotted in the rest frame of the radio galaxy assuming $z = 1.552$. This unambiguous redshift was deduced from numerous absorption lines and two strong spectral breaks (at rest wavelength $\lambda_{\text{rest}} = 2,635$ and $2,897 \text{ Å}$), features whose existence in the near-ultraviolet spectrum of the Sun and other low-mass dwarfs is long-established¹⁹, and which are evident in the ultraviolet spectra of low-redshift ellipticals such as M32 (Fig. 1c). This is, to our knowledge, the first time that such a high galaxy redshift has been derived successfully from late-type absorption features; indeed, with $V \approx 26 \text{ mag}$, this is probably the faintest galaxy for which an absorption-line redshift has ever been determined.

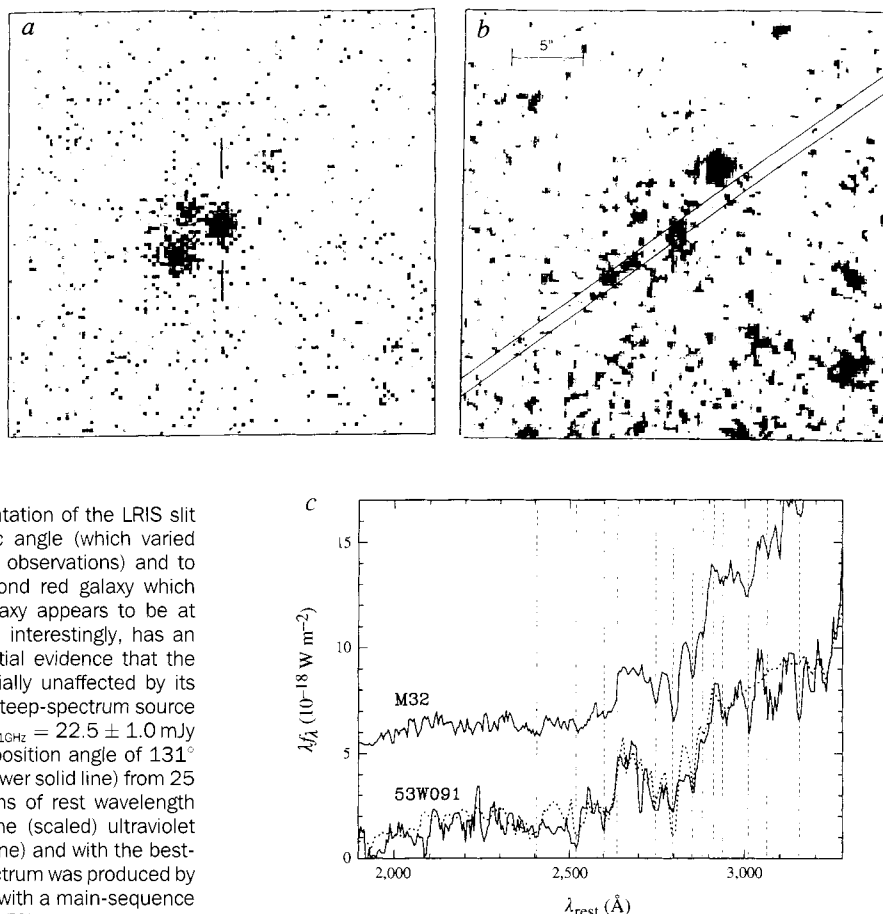
The strength of such stellar features combined with the lack of detectable emission lines (Mg II is seen in absorption, in contrast to the situation in 3C65 (ref. 9)) indicates that the ultraviolet–optical light from radio galaxies selected at millijansky flux densities is essentially free from the contaminating effects (either direct or indirect) of their active nuclei. Moreover, the similarity of this near-ultraviolet spectrum to that of low-mass main-sequence stars suggest that the red optical–infrared colour of this galaxy results from an evolved stellar population rather than, for example, a significant contribution from a dust-obscured quasar^{20–22} (a viewpoint supported by the very similar appearance of 53W091 at optical and infrared wavelengths; Fig. 1).

To investigate the extent to which our data can constrain the age of 53W091, we first calculated the best-fit age produced by an updated version of the evolutionary synthesis models of Guiderdoni and Rocca-Volmerange²³. Considering a model in which 53W091 is formed in a single burst of star-formation and evolves passively thereafter, we find that the observed optical spectrum and the infrared colours ($R - K = 5.8$, $J - K = 1.7$, $H - K = 0.8$) are all perfectly reproduced by the models only at a time of 4 Gyr after cessation of star-formation activity. Next, we considered the most recent versions of the models of Bruzual and Charlot²⁴. It is known that red optical–infrared colours are produced more rapidly by these models, and indeed we found that at $z = 1.55$ the observed optical–infrared colour is reproduced after only 1.5 Gyr. But the dependence of $R - K$ colour on age is controversial (see below) and so it was with interest that we found these same models could not reproduce the shape of the Keck spectrum until an age $> 3 \text{ Gyr}$, while to produce spectral breaks at $\lambda_{\text{rest}} = 2,635$ and $2,897 \text{ Å}$ of the strength observed in 53W091 requires an age $> 4 \text{ Gyr}$. Third, we considered the models of Bressan, Chiosi and Fagotto²⁵, which, assuming solar metallicity, yielded a best-fit age of 3 Gyr (again in good agreement with the above results).

We derived ages using these three alternative models of galaxy evolution because it is well known that different models can produce significantly different ages from a given set of data²⁶. But much of the difference between these models lies in different treatments of post-main-sequence evolution, and although this is expected to have a significant effect on the predicted infrared–optical luminosity of the galaxy, its effect on the predicted near-ultraviolet spectral energy distribution (SED) should be minimal (an expectation apparently borne out by the fact that all three models are in good agreement over the age required to reproduce the ultraviolet SED). Indeed, because it is well documented that the near-ultraviolet spectrum ($\lambda < 3,000 \text{ Å}$) of a stellar population with an age of a few Gyr should be determined simply by the

FIG. 1 Imaging and spectroscopy of 53W091. a, An infrared image of 53W091 produced by combining the IRCAM3 J- and H-band images. The field is 30-arcsec square and the radio galaxy identification is indicated by two bars. The position of this object is coincident with the centroid of the radio source (right ascension 17 h 21 min 17.81 s, declination +50° 08' 47.5" (1950)) to within 0.5 arcsec.

b, A 10-minute R-band image of the same field taken with the 10-m Keck Telescope, with lines superimposed to indicate the position and orientation of the LRIS slit which was used to obtain the optical spectrum. The position of the optical identification is coincident, to within the errors, with that of the near-infrared identification (in contrast to the highly wavelength-dependent morphologies often displayed by more powerful high-redshift galaxies such as 3C324⁸). In this 1-arcsec seeing optical image the radio galaxy is resolved, with a deconvolved full-width at half-maximum of 1.3 arcsec. The orientation of the LRIS slit (126°) was chosen to be close to the parallactic angle (which varied between 100° and 150° during the spectroscopic observations) and to enable a spectrum to also be obtained of a second red galaxy which lies a few arcsec southeast of 53W091. This galaxy appears to be at the same redshift as 53W091, ($z = 1.55$), and, interestingly, has an almost identical SED, providing further circumstantial evidence that the optical-infrared properties of 53W091 are essentially unaffected by its active nucleus. At radio wavelengths, 53W091 is a steep-spectrum source ($\alpha_{1.41\text{GHz}}^{0.61\text{GHz}} = 1.3 \pm 0.13$, where $f_\nu \propto \nu^{-\alpha}$) with $S_{1.41\text{GHz}} = 22.5 \pm 1.0$ mJy (ref. 17), and is extended (by ~ 4 arcsec) along a position angle of 131° (ref. 16). c, The 5.5-h Keck spectrum of 53W091 (lower solid line) from 25 July, 31 August and 1 September, plotted in terms of rest wavelength (assuming $z = 1.552$) and compared both with the (scaled) ultraviolet spectrum of the nearby elliptical M32 (upper solid line) and with the best-fitting model spectrum (dotted line). (The model spectrum was produced by synthesizing the spectrum produced by a Scalo IMF with a main-sequence turn-off mass of $1.35M_\odot$ (equivalent to spectral type F2), corresponding to an age of 3.5 Gyr.) The data for M32 and 53W091 have been rebinned to the same (5 Å) resolution, with an additional median smooth being applied to the latter. For ease of comparison, the spectrum of M32 has been scaled to the same amplitude as that of 53W091 at $\lambda_{\text{rest}} = 2.897$ Å and then offset vertically by $5 \times 10^{-18} \text{ W m}^{-2}$. The unambiguous nature of the redshift is demonstrated by the existence of at least 11 absorption features and two strong spectral breaks in the spectrum of 53W091 (indicated by vertical dashed lines) which are all reproduced in the rest-frame spectrum of M32 (in particular the 'top-hat' feature between 2,640 and 2,750 Å is a feature unique to this spectral range, and rules out the (remote) possibility that the break observed at 7,400 Å could be the 4,000 Å break at lower redshift).



These features (all except the reddest three of which are also reproduced in the model spectrum) are essentially those which are seen in the near-ultraviolet spectra of F and G stars, as can be judged by comparison with the near-ultraviolet spectrum of the Sun¹⁹, vindicating our belief that the red colours of the millijansky radio galaxies are a result of their evolved stellar populations. Such features have not been detected before in the spectrum of a high-redshift galaxy, not even in the Keck spectrum of the reddest powerful radio galaxy at $z \sim 1$, 3C65, in which they are apparently swamped by broad Mg II emission¹¹.

turn-off point of the main sequence²⁷, and as the main-sequence lifetime of A $\rightarrow$ G stars is probably the best understood area of stellar evolution, it should be possible to date 53W091 in an appealingly model-independent manner by simply determining the spectral type of the main-sequence turn-off point. We have therefore used the latest stellar atmosphere models of Kurucz²⁸ to investigate the spectra produced by main-sequence stars at a variety of ages, and find that the single stellar spectral type which best describes the near-ultraviolet light from 53W091 between $\lambda \approx 2,000$ and $3,500$ Å is F5 (with an effective temperature $T_{\text{eff}} = 6,500$ K). Furthermore, an independent comparison with the International Ultraviolet Explorer satellite spectra of stars of various spectral types produces exactly the same result. However, to set a realistic limit on the age of this galaxy, one must integrate over the initial mass function (IMF) of stars from very low masses ($\sim 0.1M_\odot$) up to the stellar mass at which the synthesized spectrum becomes unacceptably blue. Assuming a Scalo IMF²⁹, we find that the best-fitting main-sequence turn-off point occurs at spectral type F2 ($T_{\text{eff}} = 6,900$ K), equivalent to a stellar mass of $1.35M_\odot$. The main-sequence lifetime of a star of this mass is well established (to within 5%) to be 3.5 Gyr. Such an age is reassuringly consistent with the ages indicated by the different evolutionary synthesis codes considered above, and the fact that

this simple model provides such an excellent description of the data confirms that we are justified in ignoring the contribution of post-main-sequence stars (Fig. 1c).

We have considered carefully the robustness of our result, paying particular attention to the ways in which we could possibly have over-estimated the age of 53W091. First, if the validity of the models at near-ultraviolet wavelengths is accepted, an age younger than 3 Gyr is strongly excluded by the overall shape of the ultra-violet SED of 53W091 (Fig. 2a). Of course, the inferred age may in principle be reduced by truncating the stellar IMF at almost exactly $1.35M_\odot$, but this requires considerable fine tuning (any significant population of A stars will dominate the spectrum for ~ 3 Gyr) and, if true, it would be expected that the galaxy would display other signs of youth, such as strong emission lines which are not seen.

Second, an independent check of our derived age is provided by comparing the ultraviolet properties of 53W091 with those of nearby early-type galaxies whose ages have been determined by other means. In Fig. 2b the normalized rest-frame ultraviolet SED of 53W091 is compared with the IUE spectra of M32 and NGC2681³⁰. Recent studies of M32 both at optical³¹ and ultraviolet²⁷ wavelengths agree that the most recent star-formation activity in this galaxy occurred 4–5 Gyr ago, whereas in NGC2681 star

formation appears to have ceased 1–2 Gyr before the epoch of observation³². The level and shape of the rest-frame ultraviolet SED of 53W091 is undoubtedly more like that of M32, again consistent with an age of 3–4 Gyr for this high-redshift galaxy.

A third concern is that in comparing the ultraviolet SED of 53W091 both with models and with nearby ellipticals we may have over-estimated the age of 53W091 by ignoring the possible reddening effect of dust. But if this were the case, then it would be expected that reddening-independent features such as the strengths of the spectral breaks at $\lambda = 2,635 \text{ \AA}$, $2,897 \text{ \AA}$ would indicate a younger age. In fact these two features in the spectrum of 53W091 appear stronger than the corresponding breaks in the ultraviolet spectrum of M32, and when compared with the predictions of the main-sequence model, each break independently yields an age $> 3.5 \text{ Gyr}$ (Fig. 2c). This impressive agreement leads us to conclude that although a dusty torus may be present in the centre of this galaxy (obscuring the active nucleus) its effect on the integrated starlight of the galaxy is negligible.

The only remaining issue is the sensitivity of the derived age to the assumed value of metallicity. We have therefore investigated the effect of varying the metallicity both in our own main-sequence models, and in the full spectral synthesis code of Bressan. Both models yield very similar results; the derived age

increases to 4.5 Gyr if 1/5 of solar metallicity is assumed, but is reduced by 0.5 Gyr (to 3 Gyr) if the metallicity is doubled to twice solar (Fig. 2c). Thus, although the age–metallicity degeneracy is hard to break (similar results are obtained by fitting either to the overall SED or to the strength of the two breaks), we conclude that very large values of metallicity are required to reduce the derived age of 53W091 to less than 3 Gyr.

Having considered the ways in which we might have over-estimated the age of 53W091, we should stress that we regard it as more plausible that 3.5 Gyr may be a significant under-estimate of the true age of this galaxy. First, the spectral breaks indicate a larger age. Second, and more importantly, all of the model fits discussed above assume that this large elliptical galaxy was formed in an instantaneous burst after which star-formation completely ceased; if one assumes a burst of significant duration, or subsequent star-formation activity the derived age increases accordingly (for a 1-Gyr formation starburst, the best-fit age is 4.5 Gyr even if solar metallicity is assumed).

We now consider briefly the far-reaching implications of a 3.5-Gyr-old galaxy at $z = 1.55$. This is the first time that such an unambiguously old object has been discovered at such large look-back times, and its existence sets strong constraints both on the first epoch of galaxy formation and on cosmological models. The

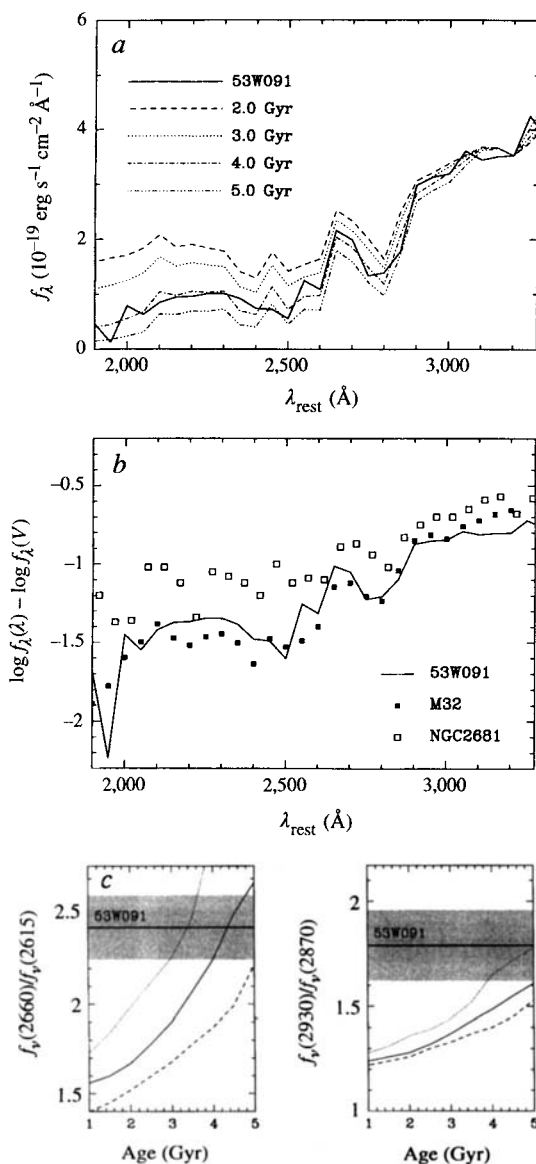


FIG. 2 Constraints on the age of 53W091. *a*, The spectrum of 53W091 (solid line), binned to a resolution of $50 \text{ \AA}$, plotted in terms of f_λ , and compared with the SED produced by the passively evolving main-sequence model at ages of 2, 3, 4 and 5 Gyr after cessation of star-formation activity. The model spectra have been normalized to the observed flux density at $3,200 \text{ \AA}$ to illustrate why ages $\leq 3 \text{ Gyr}$ can be formally rejected at a high level of significance on the basis of the overall shape of the ultraviolet SED. Models younger than 3 Gyr overpredict the flux density at $\lambda \approx 2,100 \text{ \AA}$ by $\sim 100\%$ (the relative flux-calibration of the spectrum is accurate to $\sim 10\text{--}15\%$ across the full wavelength range). We stress that, in the regime of interest here, the connection between the spectral type of the main-sequence turn-off point and age is well-established (to within 5%), because stellar ages for main-sequence stars in the mass range $1\text{--}1.5 M_\odot$ are essentially unaffected by uncertainties such as assumed mass loss, choice of convection theory or equation of state, and the opacities in the corresponding temperature range are well known. *b*, The spectrum of 53W091 (solid line), binned to a resolution of $50 \text{ \AA}$, and this time plotted in terms of $\log f_\lambda(\lambda) - \log f_\lambda(V)$ to allow direct comparison with the normalized ultraviolet SEDs of two nearby early-type galaxies, M32 (filled squares) and NGC2681 (open squares)³⁰. The rest-frame V-band flux density, $f_\lambda(V)$, of 53W091 was determined from a weighted average of the observed J- and H-band flux densities, measured through an aperture equivalent to that used in the flux calibration of the optical spectrum. Recent studies of M32 both at optical³¹ and ultraviolet²⁷ wavelengths are in agreement that the most recent star-formation activity in this galaxy occurred 4–5 Gyr ago, whereas in NGC2681 star formation appears to have ceased 1–2 Gyr before the epoch of observation³². The level and shape of the rest-frame ultraviolet SED of 53W091 is more like that of M32, consistent with an age of 3–4 Gyr for this high-redshift galaxy. *c*, The observed strength of the spectral breaks at $2,635 \text{ \AA}$ (left-hand plot) and $2,897 \text{ \AA}$ (right-hand plot) compared with the break strengths predicted by the evolving main-sequence model as a function of age for three different choices of metallicity—solar (solid line), twice solar (dotted line) and $0.2 \times$ solar (dashed line). In each plot the horizontal line indicates the break strength as measured from the spectrum shown in Fig. 1c, and the shaded area indicates the uncertainty in this estimate. The strength of the breaks in both the data and the models was determined from the ratio of the average value of f_λ in $30 \text{ \AA}$ bins centred on the two wavelengths indicated on the vertical axis of each plot. Based on the (reddening-independent) strength of these features the inferred age of 53W091 is $> 4 \text{ Gyr}$, and it is hard to reconcile the data with an age younger than 3 Gyr even if one assumes super-solar metallicity. The observed strengths of the corresponding breaks in the observed spectrum of M32 are 1.6 and 2.1 respectively, consistent with an age of $\sim 4 \text{ Gyr}$.

age of the Universe at this epoch is $1.6h^{-1}$ Gyr if $\Omega = 1$, increasing to $2.7h^{-1}$ Gyr for $\Omega = 0.2$, or at most $3.5h^{-1}$ Gyr for $\Omega = 0.2$, $\Lambda = 0.8$ ($h \equiv H_0/100 \text{ km s}^{-1} \text{ Mpc}^{-1}$). Obviously the Einstein-de Sitter model is in difficulty, and even in a low-density Universe an age >3 Gyr at $z = 1.55$ requires a formation redshift $z_f > 4$ (for

$h = 0.65$). It thus seems clear that at least some galaxies formed at redshifts greatly in excess of the recent formation era inferred from data on field galaxies³³, and the existence of similarly old galaxies at still higher redshifts would potentially allow one to infer a non-zero cosmological constant. $\square$

Received 17 January; accepted 23 April 1996.

1. McCarthy, P. J. *A. Rev. Astr. Astrophys.* **31**, 639–688 (1993).
2. Steidel, C. C., Pettini, M. & Hamilton, D. *Astr. J.* **110**, 2519–2536 (1995).
3. Steidel, C. C., Giallisco, M., Pettini, M., Dickinson, M. & Adelberger, K. L. *Astrophys. J.* **462**, L17–L20 (1996).
4. Petitjean, P., Pécontal, E., Vallis-Gabaud, D. & Charlot, S. *Nature* **380**, 411–413 (1996).
5. Dunlop, J. S., Guiderdoni, B., Rocca-Volmerange, B., Peacock, J. A. & Longair, M. S. *Mon. Not. R. astr. Soc.* **240**, 257–284 (1989).
6. di Serego Alighieri, S., Fosbury, R. A. E., Quinn, P. J. & Tadhunter, C. N. *Nature* **341**, 307–309 (1989).
7. Tadhunter, C. N., Scarrott, S. M., Draper, P. & Rolph, C. *Mon. Not. R. astr. Soc.* **256**, 53P–58P (1992).
8. Rigler, M. A., Lilly, S. J., Stockton, A., Hammer, F., Le Fèvre, O. *Astrophys. J.* **385**, 61–82 (1991).
9. Hamilton, D. *Astrophys. J.* **297**, 371–389 (1985).
10. Lilly, S. J. *Astrophys. J.* **333**, 161–167 (1988).
11. Stockton, A., Kellogg, M. & Ridgway, S. *Astrophys. J.* **443**, L69–L72 (1995).
12. Rawlings, S. & Saunders, R. *Nature* **349**, 138–140 (1991).
13. Dunlop, J. S. & Peacock, J. A. *Mon. Not. R. astr. Soc.* **263**, 936–966 (1993).
14. Windhorst, R. A., van Heerde, G. M. & Katgert, P. *Astr. Astrophys. Suppl. Ser.* **58**, 1–37 (1984).
15. Oort, M. J. A. & van Langevelde, H. J. *Astr. Astrophys. Suppl. Ser.* **71**, 25–38 (1987).
16. Oort, M. J. A., Katgert, P., Steeman, F. W. M. & Windhorst, R. A. *Astr. Astrophys.* **179**, 41–59 (1987).
17. Neuschaefer, L. W. & Windhorst, R. A. *Astrophys. J. Suppl. Ser.* **96**, 371–399 (1995).
18. Dunlop, J. S., Peacock, J. A. & Windhorst, R. A. in *Galaxies in the Young Universe* 84–87 (Springer, Berlin, 1995).

19. Morton, D. C., Spinrad, H., Bruzual, G. A. & Kurucz, R. L. *Astrophys. J.* **212**, 438–445 (1977).
20. McCarthy, P. J. *Publs. astr. Soc. Pacif.* **105**, 1051–1057 (1993).
21. Lacy, M., Rawlings, S., Eales, S. & Dunlop, J. S. *Mon. Not. R. astr. Soc.* **273**, 821–826 (1995).
22. Webster, R. L., Francis, P. J., Peterson, B. A., Drinkwater, M. J. & Masci, F. J. *Nature* **375**, 469–471 (1995).
23. Guiderdoni, B. & Rocca-Volmerange, B. *Astr. Astrophys.* **186**, 1–21 (1987).
24. Bruzual, G. B. & Charlot, S. *Astrophys. J.* **405**, 538–553 (1993).
25. Bressan, A., Chiosi, C. & Fagotto, F. *Astrophys. J. Suppl. Ser.* **94**, 63–115 (1994).
26. Charlot, S., Worthey, G. & Bressan, A. *Astrophys. J.* **457**, 626–644 (1996).
27. Magris, G. C. & Bruzual, G. A. *Astrophys. J.* **417**, 102–113 (1993).
28. Kurucz, R. *ATLAS9 Stellar Atmosphere Programs and 2km/s Grid CDROM* Vol. 13 (Smithsonian Astrophysical Observatory, Cambridge, MA, 1992).
29. Scalo, J. M. *Fund. Cosm. Phys.* **11**, 1–278 (1986).
30. Burstein, D., Bertola, F., Buson, L. M., Faber, S. M. & Lauer, T. R. *Astrophys. J.* **328**, 440–462 (1988).
31. Bressan, A., Chiosi, C. & Fagotto, F. *Astrophys. J. Suppl. Ser.* **94**, 63–115 (1994).
32. Windhorst, R. A. et al. *Astrophys. J.* **380**, 362–383 (1991).
33. Cowie, L. L., Hu, E. M. & Songaila, A. *Nature* **377**, 603–605 (1995).

ACKNOWLEDGEMENTS. We thank J. Davies for making the IRCAM3 service observations, A. Bressan for providing the age–metallicity models, and D. Burstein for supplying the IUE spectrum of M32. The United Kingdom Infrared Telescope is operated by the Royal Observatories on behalf of the UK Particle Physics and Astronomy Research Council. The W.M. Keck Observatory is a scientific partnership between the University of California and the California Institute of Technology, made possible by the generous gift of the W.M. Keck Foundation. This work was supported by the US National Science Foundation.

Evidence from a precessing pulsar orbit for a neutron-star birth kick

V. M. Kaspi*, M. Bailes†‡, R. N. Manchester†, B. W. Stappers§ & J. F. Bell¶

* IPAC/Caltech/JPL, 770 South Wilson Avenue, Pasadena, California 91125, USA

† Australia Telescope National Facility, CSIRO, Epping, NSW 2121, Australia

‡ University of Melbourne, School of Physics, Parkville, Victoria 3052, Australia

§ Mount Stromlo and Siding Spring Observatories, ANU, Private Bag, Weston Creek, ACT 2611, Australia

¶ University of Manchester, NRAL, Jodrell Bank, Macclesfield, Cheshire, SK11 9DL, UK

BIRTH 'kicks' to neutron stars, resulting from asymmetric supernova explosions, have been proposed to explain the high velocities of pulsars^{1,2}, the existence of companionless, high-velocity massive stars^{3,4}, and a putative Galactic halo of neutron stars⁵. The kick hypothesis has been controversial, because most of the evidence for kicks is indirect, and a physical mechanism to produce asymmetric explosions is as yet unknown⁶. Here we report five years of radio observations of the pulsar PSR J0045 – 7319, which is in an eccentric orbit around a B star⁷. The data show significant deviations from a simple keplerian orbit, which we interpret as arising from advance of the pulsar's periastron and spin-orbiting coupling⁸. Both effects arise because of the B star's rotationally induced equatorial bulge, however spin-orbit coupling requires the B star's spin axis to be inclined with respect to the orbital angular momentum vector; we find that the inclination angle is between 25 and 41 degrees. In the likely event that the angular momenta were aligned before the supernova explosion, this misalignment provides direct evidence that the neutron star received a kick at birth.

Radio timing observations⁷ made not long after the discovery of PSR J0045 – 7319⁹ showed that the minimum mass of the pulsar

companion is 4 solar masses ($4M_{\odot}$), and optical observations revealed a 16-mag main-sequence B star near the pulsar timing position. The association was confirmed by the detection of Doppler shifts of the companion's spectral lines at the expected orbital phase¹⁰. The amplitude of the B star's radial velocity variation determines the mass ratio in the system to be 6.3 ± 1.2 . Assuming a $1.4M_{\odot}$ neutron star¹¹, the B star's mass is $M_c = (8.8 \pm 1.8)M_{\odot}$, implying an orbital inclination angle $i = 44^{\circ} \pm 5^{\circ}$. From the known system parameters, the pulsar's distance from its companion at periastron is $(3.7 \pm 0.5)R_c$, whereas at apastron, the separation is $(35 \pm 5)R_c$, where R_c is the companion radius¹⁰, is $(6.4 \pm 0.7)R_{\odot}$. Multi-frequency timing observations have set a surprisingly tight constraint on mass loss from the B star¹², ruling out a wind as a possible source of timing perturbations. A companion rotating at a velocity typical of B stars should have a substantial equatorial bulge, and hence there should be a significant quadrupole term in the gravitational potential. As pointed out by Lai *et al.* (ref. 8), classical periastron advance and precession of the orbital plane should therefore be detectable, as the orbital and spin angular momenta couple, and precess about the fixed total angular momentum vector.

The forces that govern a pulsar's dynamics, be they classical or relativistic¹³, can be studied with unparalleled precision using radio pulse arrival times, because pulsars are near-perfect clocks. A 'residual' is the time difference between an observed pulse arrival time and that predicted by the best available timing model. Residuals versus time and orbital phase for PSR J0045 – 7319, after the subtraction of a model including only astrometric, spin and keplerian orbital parameters, are shown in Fig. 1a and b, respectively. They show significant and systematic deviations from zero, indicating this simple timing model is inadequate¹⁴. Figure 1c and d show residuals after the subtraction of a timing model including astrometric, spin (P , $\dot{P}$, $\ddot{P}$, where P is the spin period), keplerian orbital parameters, as well as a time-variable longitude of periastron ω , projected semimajor axis, $\dot{x} \equiv d(a \sin i)/dt$, and orbital period $\dot{P}_b$ (see Table 1). The significant values of $\dot{\omega}$ and $\dot{x}$ are expected from a rotationally induced quadrupole, where $\dot{x}$ is interpreted as a time-variable inclination angle, $d(i)/dt$, with the semimajor axis a constant. The reduced χ^2 for the simple keplerian model is 45, whereas that for the model in Table 1 is 1.5. The F -test shows that the improvement in the fit

TABLE 1 Parameters for PSR J0045 – 7319

Right ascension, α (J2000)	00 h 45 min 35.16 $\pm$ 0.07 s
Declination, δ (J2000)	–73° 19' 03.0 $\pm$ 0.2"
Spin period, P	0.92627590497(3) s
Period derivative, $\dot{P}$	4.4632(9) $\times 10^{-15}$
Second period derivative, $\ddot{P}$	–7.1(4) $\times 10^{-25}$ s $^{-1}$
Period epoch	MJD 49144.0000
Orbital period, P_b	4421040.6(4) s
Projected semimajor axis, $x \equiv a \sin i$	174.2576(7) light s
Longitude of periastron, ω	115.2540(5)°
Eccentricity, e	0.807949(3)
Periastron epoch	MJD 49169.21361(3)
$\dot{\omega}$	0.0259(5)° yr $^{-1}$
$\dot{x}$	1.17(2) $\times 10^{-9}$ light s s $^{-1}$
$\dot{P}_b$	–3.03(9) $\times 10^{-7}$
$\dot{e}$	< 2.2 $\times 10^{-13}$ s $^{-1}$

Best-fit astrometric, spin and orbital parameters for PSR J0045 – 7319. To minimize contamination of parameters by timing noise, the astrometric and orbital parameters were determined while fitting for a sufficient numbers of spin period derivatives to render the residuals gaussian-distributed (see, for example, ref. 26). Those parameters were subsequently held fixed when the spin parameters were fitted. The integrated electron column density was held fixed at 105.08 pc cm $^{-3}$ throughout the analysis¹², and $\dot{e}$ was fixed at zero while determining all other parameters. The parameter uncertainties in the table are our best estimates of the 68% confidence limits determined using a bootstrap Monte Carlo technique²⁷. The upper limit for $\dot{e}$ is at the 99% confidence level. Experience has shown that standard analytical formulae (for example, ref. 28) used in estimating pulse arrival time uncertainties can underestimate the true uncertainty by a factor of 2–3. Estimating uncertainties using the scatter in neighbouring groups of arrival times is impractical for PSR J0045 – 7319. To handle this difficulty, we work under the assumption that the reduced χ^2 of the gaussian-distributed residuals obtained while fitting out many spin period derivatives is unity. The F -test shows that the improvement in the fit after adding the three extra parameters is very significant; this statement holds even if the above assumption regarding the reduced χ^2 is relaxed. The fitted value for $\dot{P}$ is consistent with that expected on the basis of the known empirical relationship between $\dot{P}$ and $\dot{\dot{P}}$ (ref. 16).

after the addition of the three extra parameters is extremely significant.

The absence of systematic trends in the residuals, the low r.m.s. residual, and the significant drop in χ^2 suggest strongly that this timing model is correct. Indeed, the pulsar timing position given in Table 1 has an uncertainty reduced by a factor of four relative to that previously reported⁷ while having moved closer to the optical companion position by a factor of three. Subdividing the arrival times into four year-long segments and performing local fits of the keplerian parameters while holding $\dot{x}$, $\dot{\omega}$ and $\dot{P}_b$ fixed at zero confirms the tabulated results. The measured value for $\dot{\omega}$ is about six times larger than the value expected from general relativistic precession. It is also somewhat larger than the upper limit previously reported⁷; the longer data span and the simultaneous fit for other varying orbital parameters account for the discrepancy.

The measured $\dot{P}$ is much larger than that predicted from simple magnetic dipole braking¹⁵ and hence is probably not deterministic. Rather, it is most likely a manifestation of 'timing noise' seen in most other pulsars having characteristics similar to those of PSR J0045 – 7319¹⁶. The slightly non-gaussian quality of the residuals in Fig. 1c (and hence the deviation of the reduced χ^2 from unity) is also likely to be a remnant of timing noise.

The observed values of $\dot{\omega}$ and di/dt can be used to constrain θ , the angle between the B-star rotational and orbital angular momenta⁸. The rotationally induced quadrupole model demands a precession of the longitude of periastron given by^{8,17}

$$\dot{\omega} = \Omega_{\text{orb}} \frac{k R_c^2 \hat{\Omega}_s^2}{a^2 (1 - e^2)^2} \left(1 - \frac{3}{2} \sin^2 \theta \right) \quad (1)$$

and a precession of the orbital inclination angle

$$\frac{di}{dt} = \Omega_{\text{orb}} \frac{k R_c^2 \hat{\Omega}_s^2}{a^2 (1 - e^2)^2} \sin \theta \cos \theta \sin \Phi \quad (2)$$

where $\Omega_{\text{orb}} \equiv 2\pi/P_b$, $\hat{\Omega}_s \equiv \Omega_s/(GM_c/R_c^3)^{1/2}$ (Ω_s is the companion angular velocity), k is the apsidal constant (see, for example, ref. 18) and Φ is the unknown orbital plane precession phase.

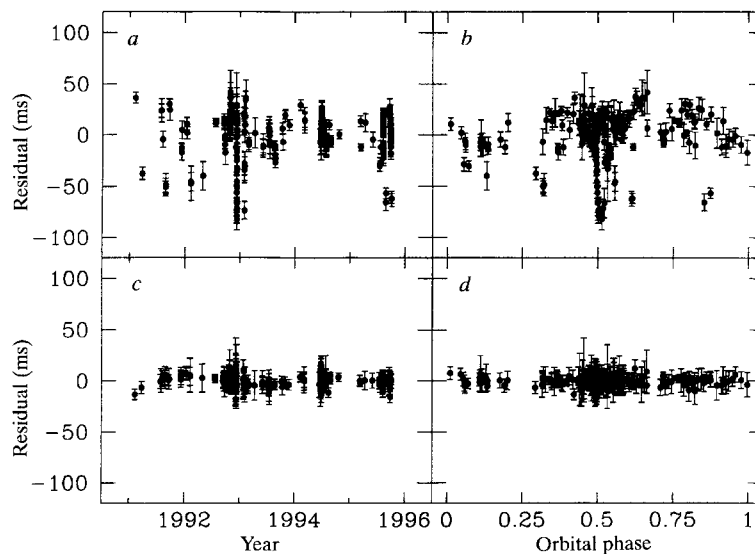


FIG. 1 Timing residuals for PSR J0045 – 7319. a, Residuals after the subtraction of a model including astrometric, spin and simple keplerian orbital parameters only. The r.m.s. residual is 19 ms, or 0.026 times the pulse period, much larger than our uncertainties. Inspection of neighbouring residuals reveals systematic, rather than random, variations. b, The same residuals as in a but plotted versus orbital phase. c, Residuals after the subtraction of the best-fit model parameters given in Table 1. The r.m.s. residual here is 3.7 ms, or 0.004 times the pulse period. d, The same

residuals as in c but plotted versus orbital phase. These data were obtained at the 64-m Parkes radio telescope in New South Wales, Australia. The observing system, as well as of our off-line data processing procedures, is described elsewhere¹². This analysis includes a total of 231 arrival times obtained at 430 MHz, 660 MHz and 1,520 MHz obtained between February 1991 and October 1995. A modified version of the standard pulsar timing analysis software package TEMPO²⁹ was used, along with the JPL DE200 ephemeris.

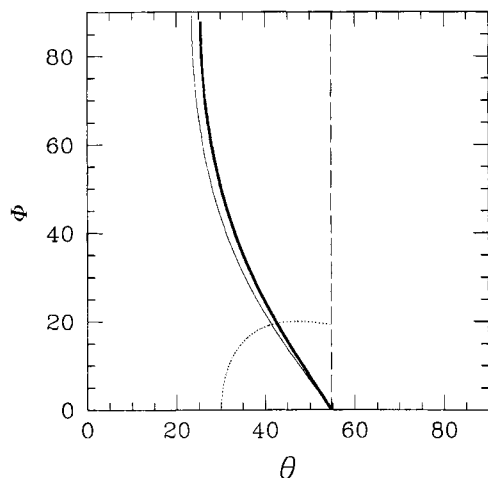


FIG. 2 Observational constraints on the geometry of the PSR J0045 – 7319 binary system. The angle θ is the angle between the B star and orbital angular momenta. The angle ϕ is the unknown orbital plane precessional phase. The region to the right of the vertical dashed line is ruled out using equation (1) because $\dot{\omega} > 0$. The heavy solid curve is the relation between θ and ϕ obtained from equations (1) and (2) and the observed ratio $\dot{\omega}/(di/dt)$, and implies that the lower limit $\theta > 25^\circ$. A correction for the expected general relativistic contribution to $\dot{\omega}$ of $0.004^\circ \text{ yr}^{-1}$ was made before the ratio was taken. The region below the dotted curve is ruled out by requiring the B star to be spinning at a velocity below the expected breakup velocity⁶ (420 km s^{-1}), and implies $\theta < 41^\circ$. Retrograde B-star rotation with respect to the pulsar's orbit is indistinguishable from prograde rotation from our data; the above angles should be subtracted from 180° in this case. A contribution to $\dot{\omega}$ of the same order as the general relativistic correction is plausible from a static tide quadrupole induced by the neutron star⁶. For a contribution equal to the general relativistic value, the lower limit rises to 29° . However, 25° is a hard lower limit corresponding to zero static tide contribution; its statistical uncertainty includes contributions from the measurement uncertainties in $\dot{\omega}$, di/dt , and the general relativistic contribution to $\dot{\omega}$. When accounting for all uncertainties, the 3σ lower limit is 23° and is indicated by a light, solid curve. The unknown true B-star breakup velocity dominates the uncertainty in the upper limit. A breakup velocity of 400 km s^{-1} gives an upper limit of 40° , whereas 500 km s^{-1} gives 43° .

Equation (1) implies $\theta < 55^\circ$, as $\dot{\omega} > 0$. The ratio $\dot{\omega}/(di/dt)$ depends on θ and ϕ only, hence the observed values for $\dot{\omega}$ and $\dot{x}$ establish a unique relation between θ and ϕ . In addition, by requiring that the B star's rotational velocity (measured¹⁰ to be $v \sin i_c = (113 \pm 10) \text{ km s}^{-1}$, where i_c is the inclination angle of the B star's spin axis to the line of sight) be smaller than the breakup velocity, $v < v_{\text{break}} \approx 420 \text{ km s}^{-1}$, a constraint on i_c , and hence on θ and ϕ is obtained⁸. Figure 2 shows these three constraints in the θ – ϕ plane. Thus the angle between the B star's spin axis and the orbital angular momentum is constrained to be

$$25^\circ < \theta < 41^\circ \quad (3)$$

In principle, equation (1) and $\dot{\omega}$ can be used to constrain k , but the large uncertainties in many of the relevant parameters preclude a precise measurement: the allowed range of parameters permits $0.0009 < k < 0.03$, and cannot constrain models¹⁸.

Before the supernova that formed the neutron star, all angular momenta in the system were likely to have been aligned, if not primordially, then because of tidal effects and mass transfer in the progenitor binary¹⁹. Any post-supernova misalignment must be a result of a kick velocity out of the original orbital plane imparted to the neutron star at the time of the supernova²⁰. Thus, the non-zero value of θ in the PSR J0045 – 7319 system provides strong new evidence that such kicks occur. Alternative misalignment

mechanisms, such as a third-body encounter, are prohibitively unlikely in environments any less dense than the cores of globular clusters. It is difficult to constrain orbital parameters of the pre-supernova binary in the case of PSR J0045 – 7319, as the progenitor system could have had a semimajor axis a_p anywhere between $a(1 - e) < a_p < a(1 + e)$, or in an even larger range if orbital evolution has taken place. For the conservative assumption of a relatively massive pre-supernova star in a wide orbit and no post-supernova orbital evolution, the kick was at least $\sim 50 \text{ km s}^{-1}$. The progenitor of PSR J0045 – 7319 is more likely to have been a closer system, because of the higher survival probability of such systems in the presence of kicks; hence the kick was probably at least 100 km s^{-1} .

The significant negative $\dot{P}_b$ implies that the orbit is decaying on a timescale, $P_b/\dot{P}_b$, of 0.5 Myr , significantly shorter than the pulsar's characteristic age, $P/2\dot{P}$, 3 Myr . Mechanisms invoked to explain orbital decay observed in some high-mass X-ray binaries (see, for example, ref. 21) are not applicable in this main-sequence-star system. Oscillation modes in the B star excited dynamically by the pulsar orbit are not likely to result in sufficient dissipation to explain the observed decay^{22,8} unless the B star's rotation is retrograde with respect to the pulsar's orbit²³. Retrograde rotation is consistent with the observations reported here, but implies an even larger birth kick velocity, indeed having a component much greater than (and directly opposite to) the pre-supernova orbital velocity. The kick velocity in this case, again for conservative assumptions, must have been at least $\sim 100 \text{ km s}^{-1}$, but could easily have been several hundred km s^{-1} . If this model is correct, the orbital evolution timescale is much shorter than the pulsar's characteristic age²³, suggesting that the latter is a poor age indicator, and controversially, that the birth spin period of the neutron star is close to its present spin period^{24,25}. Continued radio monitoring of the system to investigate the stability of $\dot{P}_b$ is planned. $\square$

Received 14 March; accepted 30 April 1996.

1. Cordes, J. M., Romani, R. W. & Lundgren, S. C. *Nature* **362**, 133–135 (1993).
2. Lyne, A. G. & Lorimer, D. R. *Nature* **369**, 127–129 (1994).
3. Gies, D. R. & Bolton, C. T. *Astrophys. J. Suppl. Ser.* **61**, 419–454 (1986).
4. Sayer, R. W., Nice, D. J. & Kaspi, V. M. *Astrophys. J.* (in the press).
5. Li, H. & Dermer, C. D. *Nature* **359**, 514–516 (1992).
6. Janka, H. T. & Mueller, E. *Astr. Astrophys.* **290**, 496–502 (1994).
7. Kaspi, V. M. et al. *Astrophys. J.* **423**, L43–L45 (1994).
8. Lai, D., Bildsten, L. & Kaspi, V. M. *Astrophys. J.* **452**, 819–824 (1995).
9. McConnell, D. et al. *Mon. Not. R. astr. Soc.* **249**, 654–657 (1991).
10. Bell, J. F., Bessell, M. S., Stappers, B. W., Bailes, M. & Kaspi, V. M. *Astrophys. J.* **447**, L117–L119 (1995).
11. Arzoumanian, Z. thesis, Princeton Univ. (1996).
12. Kaspi, V. M., Tauris, T. & Manchester, R. N. *Astrophys. J.* **459**, 717–720 (1996).
13. Taylor, J. H., Wolszczan, A., Damour, T. & Weisberg, J. M. *Nature* **355**, 132–136 (1992).
14. Kaspi, V. M., Manchester, R. N., Bailes, M. & Bell, J. F. in *Compact Stars in Binaries* (eds van Paradijs, J., van den Heuvel, E. & Kuulkers, E.) 271–277 (IAU Symp. 165, Kluwer, Dordrecht, 1996).
15. Manchester, R. N. & Taylor, J. H. *Pulsars* (Freeman, San Francisco, 1977).
16. Arzoumanian, Z., Nice, D. J., Taylor, J. H. & Thorsett, S. E. *Astrophys. J.* **422**, 671–680 (1994).
17. Smarr, L. L. & Blandford, R. *Astrophys. J.* **207**, 574–588 (1976).
18. Claret, A. & Gimenez, A. *Astr. Astrophys. Suppl. Ser.* **87**, 507–512 (1991).
19. Bhattacharya, D. & van den Heuvel, E. P. J. *Phys. Rep.* **203**, 1–124 (1991).
20. Cordes, J. M. & Wasserman, I. *Astrophys. J.* **279**, 798–802 (1984).
21. Levine, A., Rappaport, S., Deeter, J. E., Boynton, P. E. & Nagase, F. *Astrophys. J.* **410**, 328–341 (1993).
22. Kumar, P., Ao, C. O. & Quataert, E. J. *Astrophys. J.* **449**, 294–309 (1995).
23. Lai, D. *Astrophys. J.* (in the press).
24. Vivekanand, M. & Narayan, R. J. *Astrophys. Astr.* **2**, 315–337 (1981).
25. Lorimer, D. R., Bailes, M., Dewey, R. J. & Harrison, P. A. *Mon. Not. R. astr. Soc.* **263**, 403–415 (1993).
26. Kaspi, V. M., Taylor, J. H. & Ryba, M. *Astrophys. J.* **428**, 713–728 (1994).
27. Press, W. H., Flannery, B. P., Teukolsky, S. A. & Vetterling, W. T. *Numerical Recipes: The Art of Scientific Computing* (Cambridge Univ. Press, 1989).
28. Downs, G. S. & Reichley, P. E. *Astrophys. J. Suppl. Ser.* **53**, 169–240 (1983).
29. Taylor, J. H. & Weisberg, J. M. *Astrophys. J.* **345**, 434–450 (1989).

ACKNOWLEDGEMENTS. We thank D. Lai and T. Tauris for helpful discussions. Parts of the observing systems were constructed at the University of Manchester, Jodrell Bank, and the Istituto di Radioastronomia del CNR, Bologna. This research was carried out at the Jet Propulsion Laboratory, California Institute of Technology, under contract to NASA. V.M.K. received support from a Hubble Fellowship from the Space Telescope Science Institute, operated by the Association of Universities for Research in Astronomy, Inc., under contract to NASA. B.W.S. received support from an ANU Ph.D. scholarship and the ATNF student program.

Distribution of bedrock and alluvial channels in forested mountain drainage basins

David R. Montgomery*, Tim B. Abbe*, John M. Buffington*, N. Phil Peterson†, Kevin M. Schmidt* & Jonathan D. Stock*

* Department of Geological Sciences, University of Washington, Seattle, Washington 98195-1310, USA

† Simpson Timber Company, Shelton, Washington 98584, USA

MOUNTAIN river networks often consist of both bedrock and alluvial channels^{1–5}, the spatial distribution of which controls several fundamental geomorphological and ecological processes^{6,7}. The nature of river channels can influence the rates of river incision and landscape evolution^{1,2}, as well as the stream habitat characteristics affecting species abundance and aquatic ecosystem structure^{8–11}. Studies of the factors controlling the distribution of bedrock and alluvial channels have hitherto been limited to anthropogenic badlands¹². Here we investigate the distribution of channel types in forested mountain drainage basins, and show that the occurrence of bedrock and alluvial channels can be described by a threshold model relating local sediment transport capacity to sediment supply. In addition, we find that valley-spanning log jams create alluvial channels—hospitable to aquatic life—in what would otherwise be bedrock reaches. The formation of such jams depends critically on the stabilizing presence of logs derived from the largest trees in the riverside forests, suggesting that management strategies that allow harvesting of such trees can have a devastating influence on alluvial habitats in mountain drainage basins.

The spatial distribution of bedrock and alluvial reaches depends on the relation of local transport capacity (q_c) to the bedload sediment supply (q_s) delivered from upstream and across channel banks. A bedrock stream bed indicates a transport capacity in excess of sediment supply ($q_c > q_s$), whereas an alluvial stream bed indicates either a balance or an excess of sediment supply ($q_c \leq q_s$) (refs 13,14). A general expression for sediment transport capacity incorporating both drainage area, A , and channel gradient, S , is $q_c = kA^m S^n$, where $m, n > 0$ and k is an empirical variable¹⁵. In drainage basins with relatively uniform geology, hydrology and vegetation, the sediment supply to a channel reach can be expressed as $q_s = bA^p$, where b and p are empirical variables. Local conditions can cause substantial deviations in either q_c or q_s from basin-wide trends. The critical gradient, S_c , for $q_c = q_s$ may be expressed as

$$S_c = [(b/k)A^{p-m}]^{1/n} \quad (1)$$

in which b, k, p, m and n incorporate local transport capacity, sediment supply characteristics, geology and climate. This critical-slope model therefore predicts that bedrock channels occur where $S > S_c$, and alluvial channels occur where $S \leq S_c$. We tested this model against data from field surveys in the Satsop River basin, Washington.

The Satsop River originates in Tertiary basalts in the Olympic mountains, flowing southward over Tertiary marine sandstone and siltstone and Pleistocene till and outwash^{16,17} to join the Chehalis River and empty into the Pacific Ocean through Grays Harbor. The study area receives ~2,500 mm of annual rainfall¹⁸ and lies in the *Tsuga heterophylla* vegetation zone¹⁹. Extensive early twentieth century logging removed the prehistoric forest and industry now harvests naturally regenerated forest. As in most of the Pacific Northwest, early logging practices significantly disturbed stream channels²⁰, and aggressive programs intended to improve fish habitat removed logs from channels in the 1950s through to the 1980s (ref. 21). Extensive salvage of large cedar logs

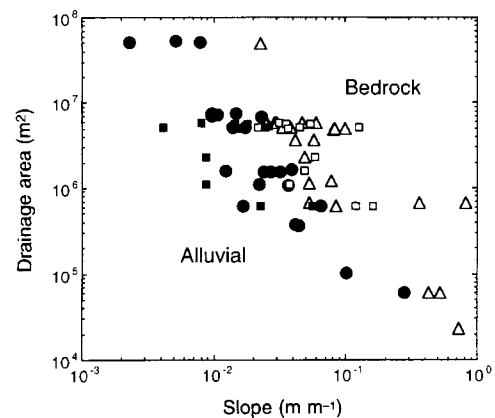


FIG. 1 Plot of drainage area versus reach-average slope for channels in the study area. Larger data points represent channel reaches lacking log jams; open triangles indicate bedrock reaches and solid circles represent alluvial reaches. Smaller data points are from reaches with channel-spanning log jams; solid squares represent forced alluvial reaches immediately upslope of valley-spanning log jams and open squares represent the slope of the underlying bedrock surface, and hence the estimated reach slope in the absence of these jams.

further diminished the supply of in-channel large woody debris over this same period.

We surveyed continuous topographic profiles down stream-bed centrelines in 59 reaches of the West Fork Satsop River and several tributaries using either a hand-level or tripod-mounted autolevel, rod and tape. We subdivided channels into bedrock or alluvial reaches ranging between 10–20 channel-widths in length, with reach-average slopes and bank full-widths of 0.0023 to 0.81 and 1 to 46 m, respectively. As considered here, bedrock reaches exhibit alluvial cover over lengths of less than a channel width, and alluvial reaches exhibit bedrock exposures of at most a channel width in length. Occasional mixed-morphology reaches with alternating bedrock and alluvial bed surfaces extending downstream for more than a channel width do occur in the study area, but were excluded from this study. We mapped bed morphology and the extent of alluvial terraces and flood plains along two tributaries onto a 1:4,000 scale map of channel thalwegs surveyed using a laser level. We also mapped reach locations onto US Geological Survey 1:24,000 scale topographic maps, from which we determined drainage areas (ranging from 0.02 to 52 km²) using a digital planimeter.

Slopes of bedrock reaches exceed those of alluvial reaches with comparable drainage areas (Fig. 1). A threshold separating bedrock and alluvial data is well described by an inverse area dependency to the critical slope, predicted by equation (1) for $m > p$. Whereas bedrock channels theoretically plot anywhere above the bedrock/alluvial threshold, non-aggrading alluvial channels would be expected to plot close to the threshold. Thus, scatter within the free-formed alluvial channels may imply that reaches with lower slopes will continue to aggrade until $S = S_c$. Alternatively, such scatter records the magnitude of local variability in sediment supply and transport capacity. Some surveyed reaches showed evidence of recent shallow landsliding from valley walls, and differences in channel confinement may impose local variability to transport capacity. Variance of the alluvial data may also reflect bed armouring and variations in roughness due to large woody debris, which can significantly reduce the transport capacity of forest channels²². The threshold shown in Fig. 1, however, indicates that such effects are a secondary influence on the spatial distribution of bedrock and alluvial reaches in the Satsop River basin.

Several processes may transiently affect the distribution of bedrock reaches in mountain drainage basins. Scour by debris flows, for example, can intermittently convert alluvial reaches to a

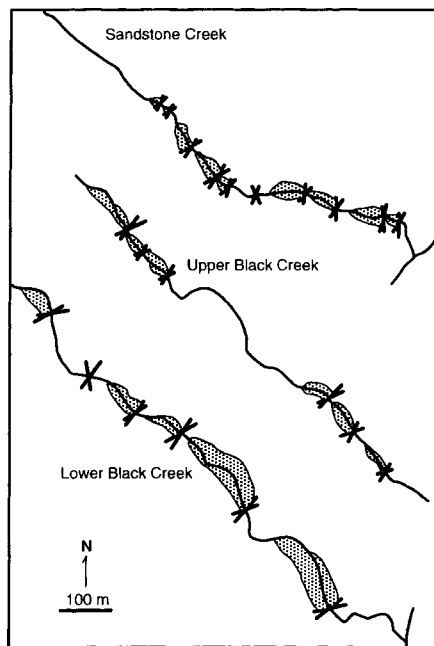


FIG. 2 Maps of southeast-flowing Black and Sandstone creeks showing the distribution of log jams (crosses) and alluvial terraces (stippled).

bedrock morphology. Episodic disturbance such as landslides may introduce and route sediment pulses through mountain channel networks, temporarily converting bedrock reaches to alluvial morphologies²³. We anticipate that spatial and temporal variations in sediment supply could complicate relations of the kind reported here; local sources of high sediment supply could result in alluvial reaches within the field of bedrock data and locally reduced supply, or increased transport capacity could likewise result in bedrock reaches within the field of alluvial data. Such exceptions to the critical-slope model may prove valuable in interpreting watershed history and conditions, a central focus of watershed analysis efforts in the Pacific Northwest²⁴.

The introduction of large woody debris can significantly influence channel morphology and aquatic ecosystems in forested environments^{25–27}. Individual logs and log jams can dominate pool and bar formation^{27–30} and control the local elevation drop in steep forest channels^{31,32}. Alluvial terraces and flood plains along surveyed tributaries of the West Fork Satsop River occur as backwater deposits immediately upstream of naturally occurring, valley-spanning log jams (Fig. 2); small terrace remnants lie upslope of both of the breached log jams that lack mappable alluvial deposits. Alluvial deposits, side channels and pools associated with these jams provide aquatic habitat unavailable in bedrock channels.

Longitudinal channel profiles indicate that physical obstruction to sediment movement and backwater effects associated with valley-spanning log jams force deposition of alluvium in otherwise bedrock reaches. In each reach with a valley jam, slopes were determined for the alluvial stream beds immediately upslope of the jam. Bedrock exposed at the base of surveyed jams and in upstream pools allowed determination of the average gradient of the underlying bedrock surface. These data reveal that log jams can force active alluvial beds in otherwise bedrock reaches (Fig. 1) and indicate that jam removal would convert these forced alluvial channels into bedrock reaches. In these mountain channels, the distribution of stable log jams dominates the distribution of alluvial channel reaches.

Individual large logs, in turn, govern the stability of valley-spanning log jams. Field observations demonstrate that the key member logs²⁷ anchoring the surveyed jams are old-growth remnants and large second-growth trees. Diameters of key members

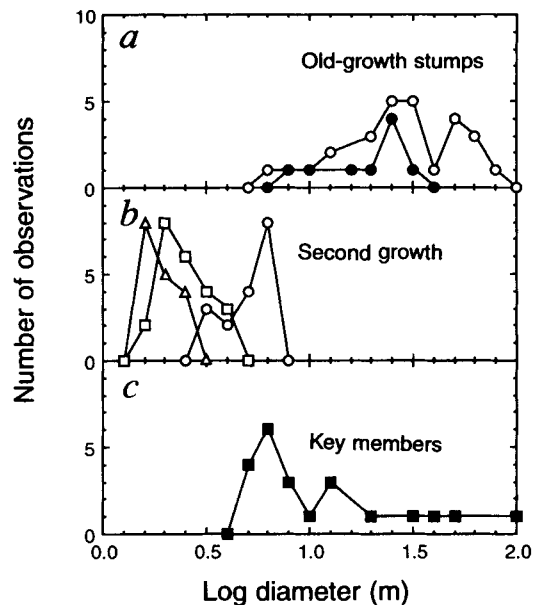


FIG. 3 Frequency distribution of a, the diameter of old-growth stumps in the riparian zone; b, trees in the current riparian forest; and c, key member logs that stabilized valley-spanning log jams. Circles, Douglas fir; filled circles, cedar; squares, hemlock; triangles, alder; filled squares, logs of any species acting to anchor jams (key members).

exceed those of all but the largest trees in second-growth riparian forests, but closely correspond to those of prehistoric riparian stumps (Fig. 3). Streamside forest management involving removal of the largest trees and retention of smaller ones^{33,34} therefore favours long-term conversion of alluvial reaches to bedrock in those portions of mountain drainage basins where $S > S_c$.

The observation that bedrock channels occur on steeper slopes than alluvial channels of comparable drainage area implies an excess transport capacity that should render bedrock channels relatively insensitive to changes in sediment supply or discharge³⁵. Reaches plotting close to the bedrock/alluvial threshold, however, likely represent channels particularly susceptible to land use impact or climate change. Channel networks with short, mixed alluvial and bedrock reaches where $S \approx S_c$ could respond dramatically to even modest changes in sediment supply, discharge or the input of large logs. Our findings support the use of a critical-slope model to discriminate bedrock and alluvial channels in landscape evolution models and suggest the potential for basin-wide predictions of channel type based on limited field surveys. Our results imply that removal of in-stream logs could have devastated much of the upland alluvial habitat in portions of mountain drainage basins in the Pacific Northwest. Furthermore, the large diameter of key member logs currently stabilizing valley-spanning jams implies that persistence of alluvial channel reaches in mountain streams can depend on the presence of old-growth-sized trees. Even the selective harvest of the largest trees in streamside forests may entail profound long-term impacts on aquatic ecosystems in forested mountain drainage basins. □

Received 26 January; accepted 8 May 1996.

1. Seidl, M. A. & Dietrich, W. E. in *Functional Geomorphology: Landform Analysis and Models* (eds Schmidt, K.-H. & de Ploey, J.) 101–124 (Catena suppl. 23, Cremlingen-Destedt, 1992).
2. Howard, A. D., Dietrich, W. E. & Seidl, M. A. *J. geophys. Res.* **99**, 13971–13986 (1994).
3. Brush, L. M. *Jr US Geol. Surv. Prof. Pap.* 282-F (1961).
4. Miller, J. R. *J. Geol.* **99**, 591–605 (1991).
5. Wohl, E. E. *J. Geol.* **101**, 749–761 (1993).
6. Richards, K. *Rivers: Form and Process in Alluvial Channels* (Methuen, New York, 1982).
7. Hynes, H. B. N. *The Ecology of Running Waters* (Univ. Toronto Press, Toronto, 1970).
8. Minshall, G. W. *Hydrobiologia* **32**, 305–339 (1968).
9. Cummins, K. W. & Lauff, G. H. *Hydrobiologia* **34**, 145–181 (1969).
10. Huryn, A. D. & Wallace, A. B. *Ecology* **68**, 1932–1942 (1987).
11. Kondolf, G. M. & Wolman, M. G. *Wat. Resour. Res.* **29**, 2275–2285 (1993).

12. Howard, A. D. & Kerby, G. *Geol. Soc. Am. Bull.* **94**, 739–752 (1983).
13. Gilbert, G. K. *Geology of the Henry Mountains* (US Geol. and Geol. Survey, Govt Print Office, Washington, 1877).
14. Gilbert, G. K. *US Geol. Surv. Prof. Pap.* 86 (Govt Print Office, Washington, 1914).
15. Kirby, M. J. in *Inst. Br. Geogr. Sp. Publ. No.* 3 (1971).
16. Logan, R. L. *Geologic Map of the South Half of the Shelton and South Half of the Copalis Beach Quadrangles, Washington* (Open File Report 87–9, Washington Div. Geol. and Earth Resources, 1987).
17. Walsh, T. J., Korosec, M. A., Phillips, W. M., Logan, R. L. & Schasse, H. W. *Geologic Map of Washington—Southwest Quadrant* (Geologic Map GM-34, Washington Div. Geol. and Earth Resources, 1987).
18. Scott, J. W., Vasquez, C. R., Newman, J. G. & Sarjean, B. C. *Washington: A Centennial Atlas* (Center for Pacific Northwest Studies, Western Washington Univ., Bellingham, 1989).
19. Franklin, J. F. & Dyrness, C. T. *Natural Vegetation of Oregon and Washington* (Oregon State Univ. Press, 1988).
20. Sedell, J. R., Leone, F. N. & Duval, W. S. in *Influences of Forest and Rangeland Management on Salmonid Fishes and Their Habitats* (ed. Meehan, W. R.) 325–368 (Am. Fish. Soc., Bethesda, 1991).
21. Sedell, J. R., Swanson, F. J. & Gregory, S. V. in *Proc. Pacific Northwest Stream Habitat Management Workshop* (ed. Hassler, T. J.) 225–245 (Humboldt State University, Arcata, 1984).
22. Buffington, J. M. thesis, Univ. of Washington, Seattle (1995).
23. Benda, L. E. thesis, Univ. of Washington, Seattle (1994).
24. Montgomery, D. R., Grant, G. E. & Sullivan, K. *Wat. Resour. Bull.* **31**, 369–386 (1995).
25. Swanson, F. J. & Lienkaemper, G. W. *USDA Forest Service Gen. Tech. Rep. PNW-69* (1978).
26. Harmon, M. E. et al. *Adv. ecol. Res.* **15**, 133–302 (1986).
27. Abbe, T. B. & Montgomery, D. R. *Reg. Riv.: Res. & Manag.* **12**, 201–221 (1996).
28. Keller, E. A. & Swanson, F. J. *Earth Surf. Proc.* **4**, 361–380 (1979).
29. Lisle, T. E. *Geol. Soc. Am. Bull.* **97**, 999–1011 (1986).
30. Montgomery, D. R., Buffington, J. M., Smith, R., Schmidt, K. M. & Pess, G. *Wat. Resour. Res.* **31**, 1097–1105 (1995).
31. Keller, E. A. & Tally, T. in *Adjustments of the Fluvial System* (eds Rhodes, D. D. & Williams, G. P.) 169–197 (Kendall/Hunt, Dubuque, 1979).
32. Marston, R. A. *Ann. Am. Assoc. Geol.* **72**, 99–108 (1982).
33. Oregon Department of Forestry *Forest Practices Water Protection Rules* ORS Ch. 527.710, OAR Ch. 629-57-2240, 629-57-2260 (1994).
34. Administrative code 222-30-020 (3) (Washington Univ., 1995).
35. Howard, A. D. in *Thresholds in Geomorphology* (eds Coates, D. R. & Vitek, J. D.) 227–258 (Allen & Unwin, London, 1980).

ACKNOWLEDGEMENTS. We thank Simpson Timber Company and the Washington State Timber/Fish/Wildlife agreement for supporting this research. Discussion with Bill Dietrich and comments by Alan Howard helped clarify our conceptual framework and greatly improved the manuscript.

CORRESPONDENCE should be addressed to D.R.M. (e-mail: dave@bigdirt.geology.washington.edu).

Free-standing and oriented mesoporous silica films grown at the air–water interface

Hong Yang*, Neil Coombs†, Igor Sokolov* & Geoffrey A. Ozin*

* Materials Chemistry Research Group, Lash Miller Chemical Laboratories, University of Toronto, 80 St George Street, Toronto, Ontario, Canada M5S 3H6

† Imagetek Analytical Imaging, 32 Manning Avenue, Toronto, Ontario, Canada M6J 2K4

SURFACTANT assemblies can function as templates for the deposition of silicates to form mesoporous silicas¹. Recently we described a surfactant-templated synthesis of oriented mesoporous silica films grown at the mica–water interface². Here we show that such films can be grown without a solid substrate, by surfactant templating at the interface between air and water. The films are continuous and have a root-mean-square surface roughness of about 3 Å. They are resilient enough to withstand significant bending, and are sufficiently flexible to be transferred onto substrates of different shapes. We propose a model for film formation which ascribes a dual-templating role to the surfactant: we suggest that both a surfactant overstructure at the air–water interface and micellar aggregates in solution interact collectively with the soluble, polymerizable silicate building blocks. These films might find applications in catalysis, separation technology and biomedicine.

Synthetic methods that enable the fabrication of oriented films of mesostructured materials provide opportunities for detailed studies of their anisotropy and size-tunable properties. This knowledge can be usefully exploited for the development of new technologies that depend on the special properties of this class of materials. Periodic mesostructures that are of current interest include semiconductors³, metal clusters⁴, porous solids⁵, composites⁶, block-copolymers⁷, liquid crystals⁸ and patterned self-assembled monolayers⁹.

Although oriented mesoporous thin films might find many applications as selective inorganic membranes, those that we described recently grown at the interface between mica and water² suffer from several limitations. In particular, few substrates have an atomically smooth surface and can sustain their structural integrity under the corrosive conditions of a synthesis, while at the same time facilitating oriented film formation. We considered that this problem might be surmounted by dispensing with a solid substrate and growing the films instead under surfactant overstructures at the interface between air and water.

The synthesis of mesoporous silica films at the air–water interface under acidic conditions¹⁰ is achieved using the following reactant mole ratios: 100 H₂O : 7 HCl : 0.11 CTACl : 0.07–0.13 TEOS; where CTACl is the cationic surfactant CH₃(CH₂)₁₅N(CH₃)₃Cl and TEOS is the silica source reagent (C₂H₅O)₄Si. The surfactant solution is mixed with TEOS and stirred for 2–3 minutes at room temperature and transferred into a polypropylene bottle. The film-forming process works well under static conditions at 80 °C over a reaction time of minutes to days. Well formed films, with thicknesses from tens of nanometres up to about half a micrometre have been grown at the air–liquid interface.

Scanning electron microscopy (SEM) images of the films that have been transferred onto a copper grid reveal that they are continuous, Fig. 1a. Polarized optical microscopy images (not shown) recorded at room temperature depict that the films are optically isotropic. This implies that the silica wall material is amorphous and that the micellar assembly contained within the channels is not behaving like a liquid crystal.

Transmission electron microscopy (TEM) images of microtomed sections cut orthogonally to the surface of the as-synthesized free-standing films show that the channels are hexagonally close-packed with a centre-to-centre distance of ~50 Å, Fig. 1b. This implies growth of the channels in an orientation parallel to the air–water interface. The observation of a smooth film surface grown at the air–water interface with a root-mean-square roughness of ~3 Å indicates that the growth process occurs via deposition of silica-surfactant micellar solution precursors at a surfactant overstructure localized at this interface. The film surface growing into the solution shows roughness on the mesoscopic scale, which might represent a silica replica of the disposition of micellar structures as they are deposited at this lower surface (Fig. 1c).

Early-stage free-standing films, lifted onto TEM grids and viewed directly, are sufficiently thin to be imaged at an accelerating voltage of 50 kV (Fig. 1d). Under these conditions, a highly ordered periodic structure with a spacing of ~25 Å is observed. The observed 25 Å periodicity is consistent with a hexagonal close-packed arrangement of one-dimensional channels viewed orthogonally to the film surface¹¹. The image is consistent with the proposal that the channels are oriented parallel to the surface of the film. TEM images (not shown) of calcined (450 °C for 4 h in air) free-standing films demonstrate that the mesopore order is intact.

SEM images show that the films are resilient to bending (Fig. 1e). They have been transferred onto substrates with different shapes. These properties may reflect the thin organic–inorganic composite nature of the films. Together with complementary powder X-ray diffraction studies, the results demonstrate that the macroscopic integrity and the mesopore order of the attached films are intact.

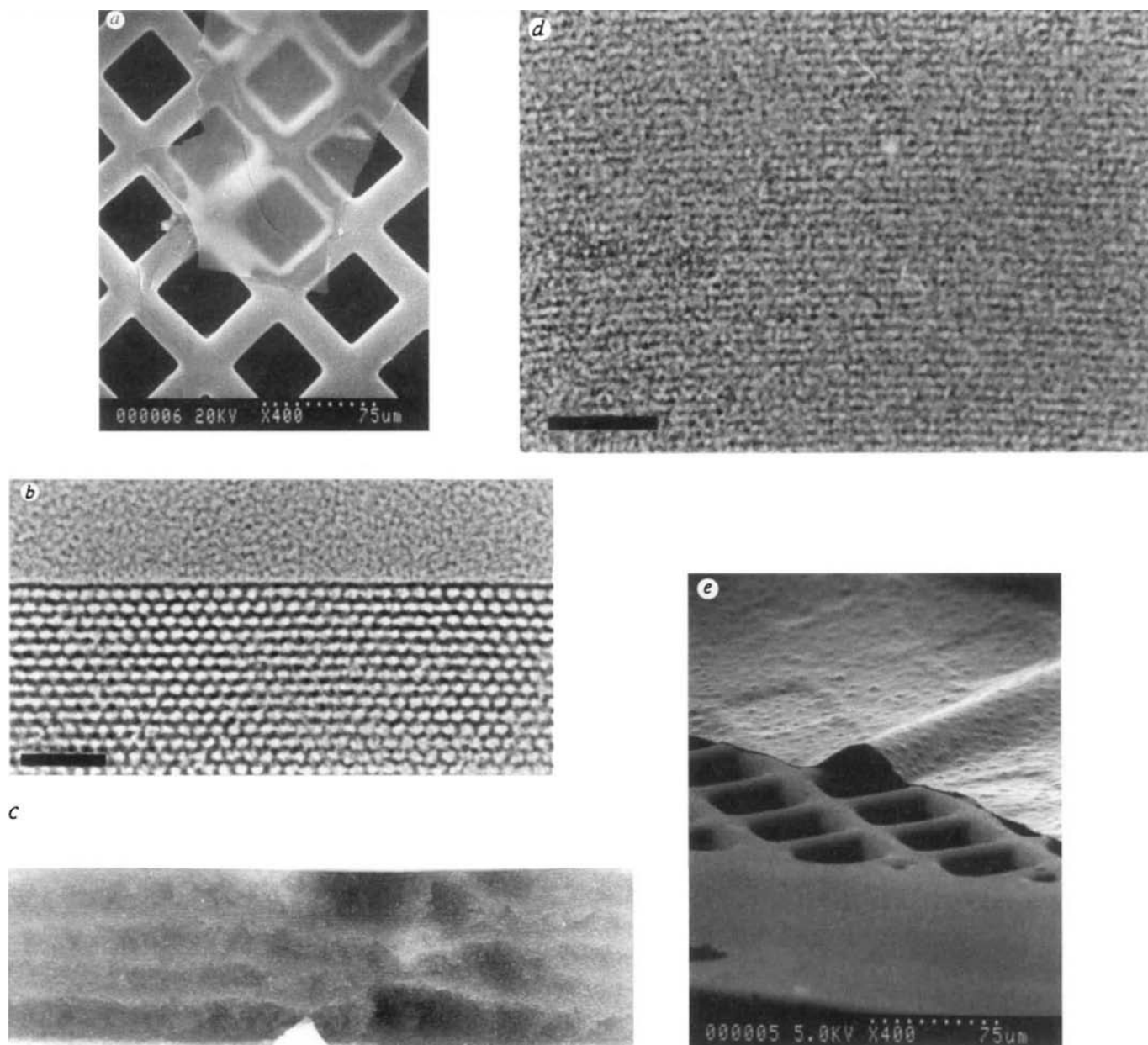


FIG. 1 *a*, SEM image (recorded on a Hitachi S-570 microscope at an accelerating voltage of 20 kV) of a free-standing film that has been transferred from the air–water interface onto a copper grid. The film is continuous, and is transparent under the imaging conditions. Dotted scale bar, 75 μm . *b*, TEM image (obtained on a Philips 430 microscope operating at 100 kV, epoxy-embedded and microtomed 100–300 $\text{\AA}$ sections cut orthogonally to the film surface) of an as-synthesized free-standing film. The film surface shown is the one formed at the air–water interface, where the channels are seen to run parallel to this interface. Scale bar, 25 nm. *c*, TEM image of an as-synthesized free-standing film (prepared for imaging as in *b*) showing the difference between the surface roughness at the top of the film (which grows at the air–water interface) and the bottom (which grows into the solution). Scale bar at the bottom, 250 nm. *d*, TEM image of a free-standing film (viewed normal to the air–water interface) showing a highly ordered periodic structure consistent with a hexagonal close-packed arrangement of channels running parallel to the surface of the film. Scale bar, 25 nm. *e*, SEM image of a free-standing mesoporous silica film which has been transferred from the air–water interface onto a copper grid. The film shows extensive regions of film curvature. Dotted scale bar, 75 μm .

Atomic force microscopy (AFM) is an effective probe of the overall topological landscape and structural detail of the surface of the mesoporous silica films (Fig. 2). AFM scans over large areas of the film provide an estimate of the root-mean-square roughness of the surface grown at the air–water interface to be $\sim 2 \text{ \AA}$ which is in good agreement with the TEM results (Fig. 1). AFM images of the films reveal parallel surface features with a periodicity of $\sim 50 \text{ \AA}$ which appear to have a spheroidal texture. This is consistent with the TEM observations of the internal channel architecture of the films. The main features in these AFM images were reproducible with respect to different scans, scan directions, load forces, tips and samples, and are not considered to portray imaging artefacts. Some of the irregularities observed in the surface of the film could arise from inhomogeneous silicate condensation processes. Overall, the observed surface topology may be viewed as a silica replica of the distribution of head groups

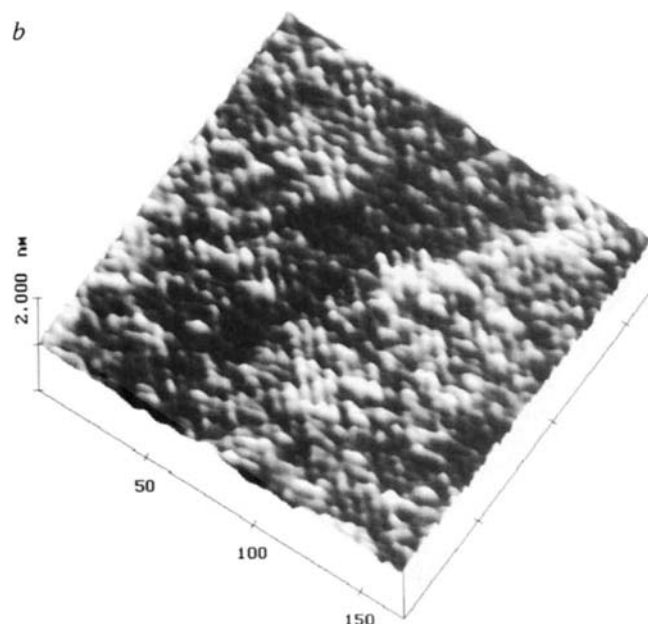
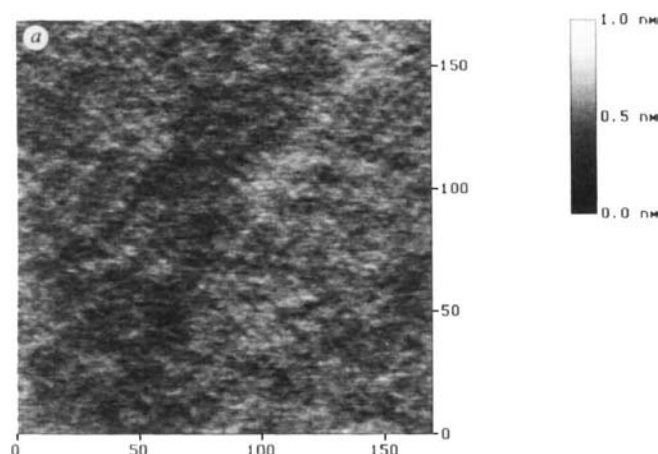


FIG. 2 *a*, Raw AFM image (using contact mode on a Nanoscope III with a silicon nitride integrated tip, Digital instruments, CA) of the surface of a mesoporous silica film, grown at the air–water interface and subsequently lifted onto a freshly-cleaved mica surface. Grey scale represents the surface height. (Units on *x* and *y* axes, nm.) *b*, Processed (filtered by a two-dimensional fast Fourier transformation) AFM image of *a* with just the high-frequency noise removed, revealing parallel surface features with a periodicity of 50 Å and a spheroidal texture. (Units on *x* axis, nm).

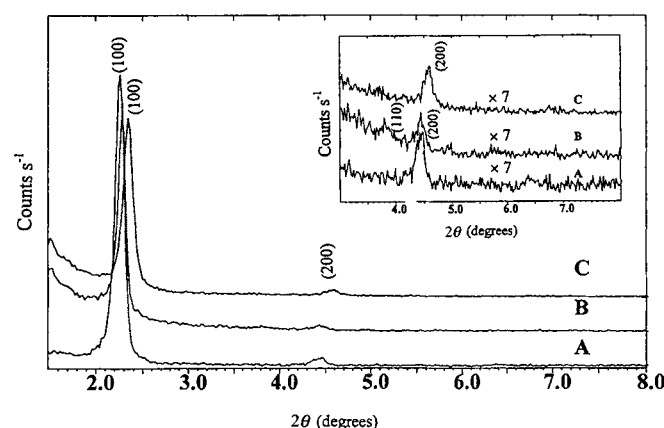


FIG. 3 Powder X-ray diffraction patterns (recorded on a Siemens D5000 diffractometer using Ni-filtered Cu $K\alpha$ radiation with wavelength $\lambda = 1.54178$ Å). Traces A and C, free-standing as-synthesized (A) and calcined (C) mesoporous silica films, showing (100) and (200) reflections. Trace B, powdered samples obtained by removing the as-synthesized films from the substrate, showing the expected (100, 110, 200) reflections. The absence of the (110) reflection for the film samples confirms that the channel axis is oriented parallel to the templating liquid surface. Inset, PXRD reflections for $2\theta = 3^\circ$ – 8° with an intensity (counts per second) scale expansion of $\times 7$.

of the surfactant overstructure and the micellar moieties existing at the interface between air and water.

The powder X-ray diffraction patterns for the as-synthesized and calcined free-standing films, lifted onto a solid substrate, are shown in Fig. 3. They both reveal (100) and (200) reflections consistent with the TEM observations that the channels run parallel to the surfactant overlayer at the air–water interface. By contrast, powdered samples obtained by removing these films from the substrate show the expected (100, 110, 200) reflections. The absence of the (110) reflection for the film samples confirms that the channel axis is oriented parallel to the templating liquid

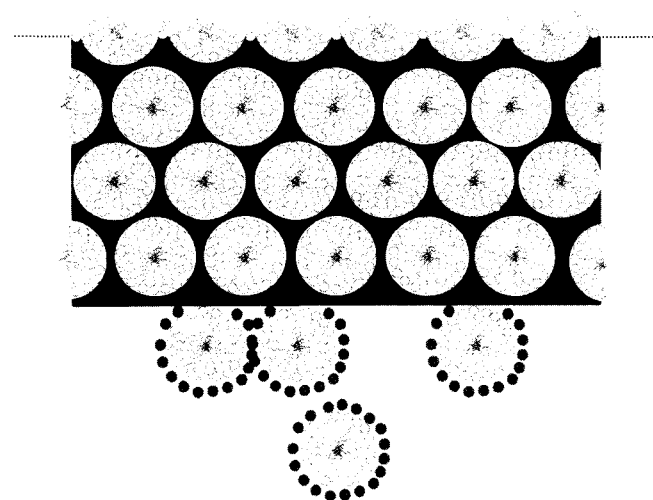


FIG. 4 Graphical illustration of the proposed model for the formation of a free-standing oriented mesoporous silica film at the air–water interface. This model involves a dual-templating role for the surfactant, and is based on collective interactions between polymerizable silicate building-blocks, a surfactant overstructure and solution micellar aggregates. Light grey, surfactant; grey, silicate building-block; black, silica. The air–water interface is represented by the dotted line at the top of the figure. The white background at the bottom of the figure represents bulk aqueous synthesis mixture.

surface, consistent with the TEM results. On calcination of the films, the anticipated contraction ($\Delta d_{100} = 2$ – 5 Å) of the hexagonal *ab*-unit cell is observed, due to the removal of the surfactant template from the channels, concomitant with the condensation of silanol (SiOH) groups in the channel walls^{1,2}.

A model for the growth of smooth and oriented mesoporous silica films at the air–water interface must take into account recent studies of mechanistic pathways for the formation of mesoporous silicas in the bulk powdered form¹² and the surface profile of soluble surfactant overlayers at the air–water interface^{13–16}. First, the general consensus is that mesoporous silicas

form through the co-assembly of surfactant-silicate micellar aggregates into a silicotropic mesophase which undergoes condensation-polymerization into a silica replica of the templating interfaces¹². Second, the arrangement of head groups for surfactants in water-soluble overlayers has recently been studied by X-ray and neutron reflectometry as well as equilibrium molecular dynamics simulation techniques^{13–16}. In the case of CTACl, surfactant overlayers are in equilibrium with the underlying solution. Published results^{13–16} suggest that the surface profile has a surfactant 'hemi-micellar' wave-like structure rather than a planar close-packed one as found for surfactants in insoluble monolayers. The surface roughness, as defined by the gaussian distribution of CTACl head groups, is $\sim 3\text{--}4\text{ \AA}$.

We propose that the formation of a mesoporous silica film involves collective interactions between silicate building-blocks, micellar solution species and a surfactant 'hemi-micellar' overstructure localized at the air–water interface (Fig. 4). Under the synthesis conditions used, film formation is considered to involve polymerization of silicates in the surfactant head group regions of a hexagonal mesophase that is concentrated at the liquid surface overstructure. This is reasonable as hexagonal phases have been observed at the liquid surface of micellar solutions¹⁶. Thus a pre-organized CTACl overlayer facilitates the organization of a smooth oriented mesoporous silica film at the air–water interface. The observed root-mean-square roughness of the films reflects the outcome of a cooperative surface process in which silicates undergo a condensation-polymerization at a templating overstructure. Film growth is probably regulated by matching charge and geometry between micellar aggregates and silicates at a surfactant structured air–water interface.

The availability of free-standing and oriented mesoporous silica films provides opportunities for tailoring their structures, pore sizes and compositions, altering their adsorption properties, and chemically modifying their surface reactivity. These mesostructured films may find uses as organized catalysts, membranes for biomolecule separations, sensors for large molecules, ordered ceramic matrix composites, engineered biomaterials and mesostructured semiconductors. □

Received 22 March; accepted 13 May 1996.

1. Kresge, C. T., Leonowicz, M. E., Roth, W. J., Vartuli, J. C. & Beck, J. S. *Nature* **359**, 710–712 (1992).
2. Yang, H., Kuperman, A., Coombs, N., Mamiche-Afara, S. & Ozin, G. A. *Nature* **379**, 703–705 (1996).
3. Alivisatos, A. P. *Science* **271**, 933–937 (1996).
4. Whetten, R. L. et al. *Adv. Mater.* **8**, 428–433 (1996).
5. Behrens, P. *Angew. Chem. int. Edn engl.* **35**, 515–518 (1996).
6. Wu, C.-G. & Bein, T. *Science* **266**, 1013–1015 (1994).
7. Leibler, L. & Fredrickson, G. H. *Chem. Br.* 42–45 (1995).
8. J. Mater. Chem. (special issue on liquid crystals) **5**, (1995).
9. Kumar, A., Abbott, N. L., Kim, E., Biebuyck, H. A. & Whitesides, G. M. *Acc. Chem. Res.* **28**, 219–226 (1995).
10. Huo, Q. et al. *Nature* **368**, 317–321 (1994).
11. Yada, M., Machida, M. & Kijima, T. *J. chem. Soc. chem. Commun.* 769–770 (1996).
12. Firouzi, A. et al. *Science* **267**, 1138–1140 (1995).
13. Bocker, J., Schlenkerich, M., Bopp, P. & Brackmann, J. *J. phys. Chem.* **96**, 9915–9922 (1992).
14. Lu, J. R., Li, Z. X., Smallwood, J., Thomas, R. K. & Penfold, J. *J. phys. Chem.* **99**, 8233–8243 (1995).
15. Thomas, R. K. & Penfold, J. *Curr. Opin. Colloid Interface Sci.* **1**, 23–33 (1996).
16. Braun, C., Lang, P. & Findenegg, G. H. *Langmuir* **11**, 764–766 (1995).

ACKNOWLEDGEMENTS. We thank S. Mamiche-Afara and A. Kuperman for assistance, and G. Henderson for the use of his AFM equipment. H.Y. thanks the University of Toronto for a Connaught Foundation graduate scholarship and a University of Toronto Open Scholarship. This work was supported by the Mobil Technology Company.

CORRESPONDENCE should be addressed to G.A.O. (e-mail gozin@alchemy.chem.utoronto.ca)

Seismicity and stress rotation in a granular model of the brittle crust

David R. Scott

Department of Geological Sciences, University College London,
Gower Street, London WC1E 6BT, UK

THE basic mechanical process responsible for earthquakes and faulting is not known. The intuitive notion of frictional slip on faults between elastic crustal blocks cannot be reconciled with laboratory measurements of the strength of rocks¹, field observations of heat flow^{2,3} and stress orientation⁴ around the San Andreas fault in California, and seismological estimates of the energy radiated by earthquakes⁵. The weakening of large faults by elevated pore-fluid pressure⁶ has been suggested as one solution to this paradox, but places severe constraints on the hydrological conditions of the faults concerned. Here I propose an alternative model for earthquake mechanics, in which the crust is treated as a system of many interlocking blocks divided by many faults⁷. The model combines this random granular structure with simple, deterministic mechanical interactions. Numerical simulations of the deformation of an aggregate of rough grains under compressive stress show earthquake-like elastodynamic failures without frictional heat production, and substantial rotation of stresses across shear zones, which mimics field observations. There remain problems of scale in comparing these simulations with nature, but a seismological test of the model may be possible.

The Earth's crust can be thought of as a cohesive elastic material divided up by a large number of faults, or equivalently as a non-cohesive granular material that is held together by the pressures due to burial. Pursuing the latter view, simulations are presented

that use the discrete element method (DEM), a relatively transparent approach to tracking forces, motions and rotations in a compressed aggregate of discrete grains (Fig. 1). This method was devised⁸ to simulate quasi-static deformation, but uses equations of motion that admit elastodynamic effects such as elastic wave propagation. Each grain has the mass and moment of inertia of a circular disk of uniform density, and is acted on by elastic forces at its points of contact with neighbouring grains. The following microscopic physical model is used to compute contact forces.

There is no force between a pair of grains unless they overlap. A repulsive central force then acts, proportional to the overlap and to an elastic modulus that is uniform for all contacts. Shear forces may also act tangentially at a contact, in a manner analogous to a pair of meshed gear wheels; there is no resistance to rolling but elastic forces resist opposing tangential motion. Previous DEM studies have placed an upper limit on shear forces, allowing intergranular slip when the ratio of shear to normal force at a contact exceeds a prescribed coefficient of friction^{8,9}. The prescribed value is typically around 0.6, corresponding to laboratory measurements on rock samples with relatively smooth and flat surfaces. Here there is no frictional limit; intergranular slip can only occur when grains separate. This is justified as an attempt to simulate a realistic material in which the contacts between grains are not smooth and flat, but rather the scale of grain irregularity is comparable to the grain size¹⁰. This literal pursuit of the gear-wheel analogy emphasises local dilation as the key process leading to intergranular slip, rather than increases in shear force.

This novel contact law has important consequences for energy dissipation. With frictional contacts, stored elastic energy is lost from the system during frictional slip, that is converted to heat. Here there is no loss of elastic energy until two grains actually separate (or unmesh). At this point any residual shear force is abruptly lost, disrupting the near-equilibrium force balance on the two grains and causing them to accelerate and rotate in opposite directions. All the elastic energy from the residual shear force flows into elastodynamic radiation. In this Letter, such slip events

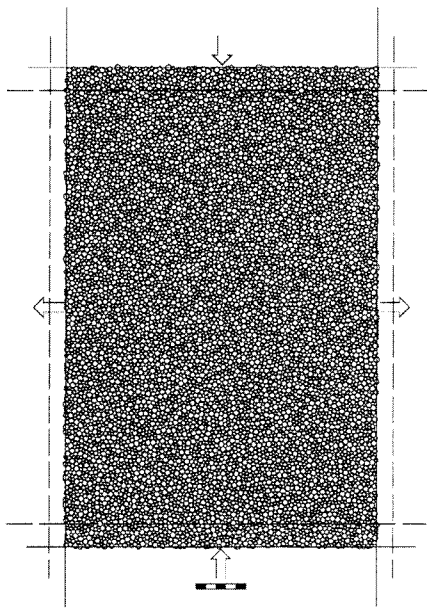


FIG. 1 The geometry used for the simulations is a two-dimensional rectangular cell that is periodic in both spatial dimensions. The cell contains an assemblage of 10^4 circular grains of three sizes; the relative radii are 3:4:5, in proportions 5:3:2 respectively. This variation in radius prevents the grains from becoming ordered onto a hexagonal lattice; the assemblage is microscopically heterogeneous, and macroscopically both homogeneous and isotropic. To prepare the assemblage, non-overlapping grains are inserted at random into a cell of larger size and compacted by reducing the size of the cell. The shear test shown in Fig. 2 is conducted by shortening the periodic cell in one dimension and extending it in the other as indicated by the arrows; the assemblage is deformed in pure shear at constant volume. The shapes of the cell at the start and end of this deformation are shown by the solid and dashed outlines, respectively. The scale bars here and in Figs 3 and 4 have lengths of ~ 13 grain diameters.

are called 'acoustic emissions' (AEs), by analogy with laboratory tests on brittle materials in which acoustic energy is used to detect internal failure¹¹.

The main difficulty in DEM is efficient detection of new contacts that form during the arbitrary rearrangement of grains. In the code used here an adaptive graph¹² records and updates the local topology of grains and contacts. The equations of motion are solved using an explicit second-order finite-difference scheme, with a constant time step chosen to resolve accurately the motion of grains with the smallest natural period of vibration⁹ (computed from the mass of a grain and the combined stiffness of its contacts). A small and uniform viscous damping force is applied to all grains, dissipating kinetic energy that would otherwise be trapped indefinitely by the periodic geometry. The resulting attenuation of elastic waves is independent of frequency¹³, and the scale distance for dissipation of wave energy is chosen to match the size of the periodic cell.

Figure 2 shows both inelastic and elastodynamic information extracted from a typical simulation in which the ratio of mean stress to the elastic modulus of the grains in the compacted assemblage is approximately 0.01. This corresponds to a typical seismogenic depth of 10–20 km into the crust. The only other significant adjustable parameter is the loading speed, which is discussed below. The simulation reproduces the main features of mechanical tests on brittle materials. The initial deformation is uniform and elastic, but becomes inelastic as the energy and rate of AE increases. The stress climbs to a peak as the strain starts to localize onto a system of conjugate oblique shear zones, and then drops as the strain is localized to a single oblique shear zone. A persistent state is reached which, in this periodic geometry, corresponds to the shearing of an infinite deck of cards. The

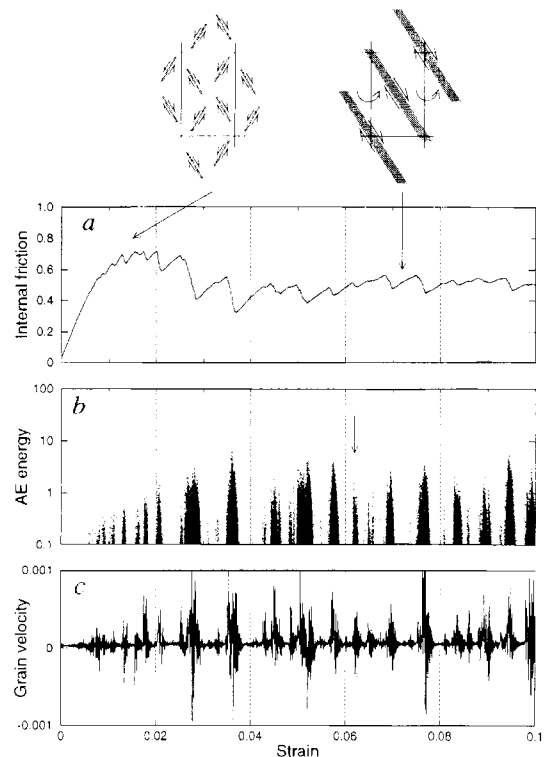


FIG. 2 Summary of results for shear deformation of the assemblage shown in Fig. 1. The horizontal axes shown strain, which is equivalent to time because the strain rate is constant. *a*, Internal friction, a measure of the ratio of shear stress to mean stress averaged over the assemblage. The sketches above the panel show the pattern of shear localization observed at two stages of the deformation (arrowed). *b*, Energy released by each acoustic emission (AE), scaled to the mean energy stored in a single elastic contact. The arrow indicates the earthquake illustrated by Fig. 3. *c*, One component of the velocity of a single grain within the sample, scaled to the speed of elastic waves in the granular assemblage. This 'seismogram' is smoothed over 100 time steps to show coherent, longer-period motions. Both the acoustic emissions and the seismogram show bursts of activity separating periods of quiescence.

development of shear zones is a ubiquitous feature of deformation in soils¹⁴ and rocks¹⁵ and has been considered in numerical studies using granular^{16,17} and continuum^{18,19} mechanics, as well as in theoretical studies^{20–22}. Two other features of the simulation invite comparison with geophysical observations.

First, the pattern of AEs mimics the basic characteristics of earthquakes. The AEs are clustered in time, each cluster representing an 'earthquake'. Figure 3 shows how the AEs in a single earthquake are localized in space and how the 'co-seismic' displacement and strain resembles that of a shear dislocation or strike-slip fault²³. The long-term inelastic strain is produced by a series of such earthquakes within the shear zone. The rupture process is elastodynamic; once an earthquake has started it cannot be stopped by stopping the external loading, and its duration is determined by the elastic wave speed. The interval between earthquakes is, however, determined by the external loading speed. The ratio of loading speed to elastic wave speed in this simulation is approximately 10^{-3} ; the equivalent ratio for the deforming crust is around 10^{-12} , so the intervals between real earthquakes are much longer relative to their duration. The loading speed in the simulation is, however, demonstrably slow enough to prevent earthquakes from blurring into one another (see Fig. 2b).

Second, although the shear zone is defined a locus of strain localization, it can also be identified in the stress field. Figure 4 shows that shear localization leads to dilation of the material within the shear zone and to a pronounced rotation of the orientation of maximum compressive stress. This rotation devel-

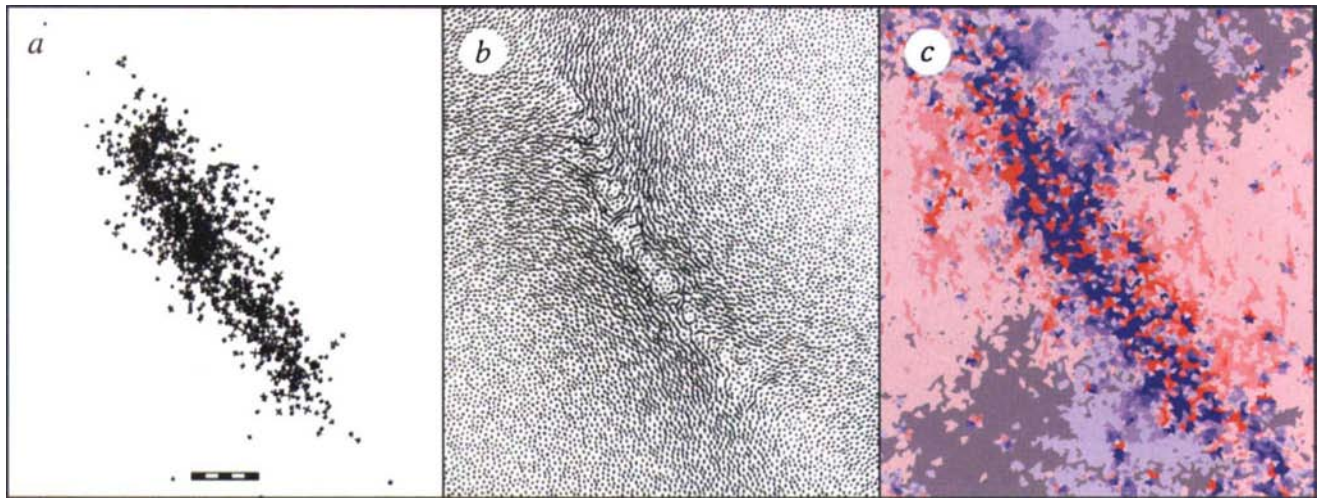
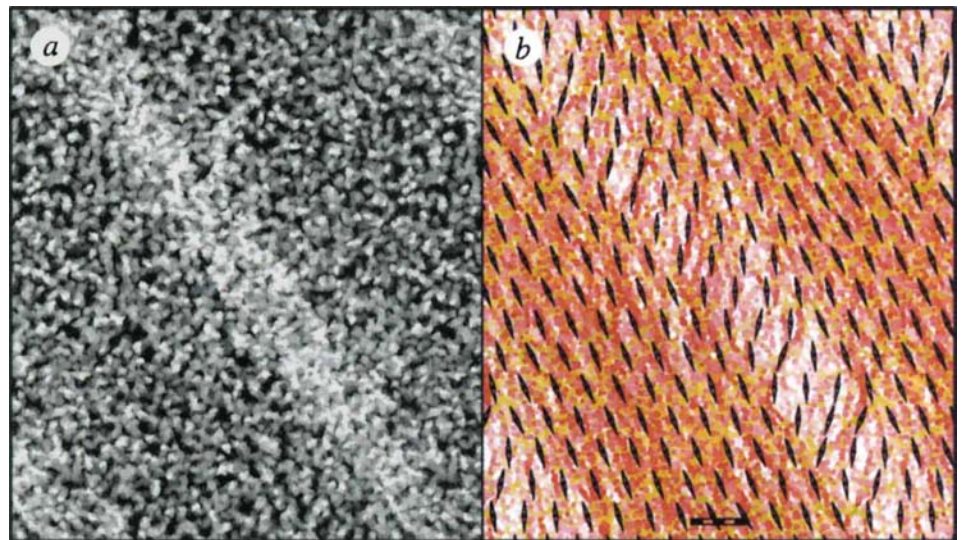


FIG. 3 'Coseismic' deformation for a time interval spanning the earthquake indicated in Fig. 2b; the behaviour shown is typical. The area shown is slightly smaller than the periodic cell. *a*, Locations of the acoustic emissions that define this earthquake. The size of each cross is proportional to the moment of the corresponding acoustic emission. *b*, Displacement vectors for each grain, defining a right-lateral strike-slip 'fault'. The displacements have been magnified by a factor of four, so the earthquake strain is approximately 1%. *c*, Volumetric strain derived from

the displacement field; red and blue correspond to compaction and dilation respectively, with the most intense colours showing strains lying outside the limits of $\pm 1\%$. The four prominent lobes correspond to the radiation pattern for compressional seismic waves from a strike-slip fault. In *b*, the two sides of the fault are separated by a zone of incoherent rotation and strain that is up to 5 grain diameters wide. This appears to be the sharpest dislocation that can be accommodated by the heterogeneous granular microstructure.

FIG. 4 Porosity and stress distributions immediately after the earthquake shown in Fig. 3; the patterns shown are typical. The area shown is slightly larger than the periodic cell. The porosity is shown in *a*; the lightest and darkest shades show porosities above and below the limits of 0.2 and 0.1 respectively. The stress distribution is shown in *b*. The colour of each grain ranges from yellow to red, indicating the ratio of shear stress to mean stress for that grain. Grains with relatively isotropic or hydrostatic loading are yellow, whereas those that are preferentially loaded in a single direction are red. In addition, the intensity of colour indicates the mean stress level, so that grains carrying relatively little load are pale. Lines of bright red grains show load-bearing bridges within the assemblage³⁶. The compass points emphasize the local orientation of maximum compressive stress, averaged over a circular region of diameter equal to the length of the pointer. The main shear zone can be identified clearly in *a* and *b*. The porosity in the shear zone is greater than of the surroundings. The stress field outside the shear zone is relatively homogeneous and shows a uniform rotation of the maximum compressive stress towards the parallel to the shear zone. The stress field inside the



shear zone is much more heterogeneous, with pale regions that carry very little load separated by red load-bearing bridges. The finite right-lateral strain of the shear zone has rotated these grain bridges, and hence the stress field, towards the perpendicular to the shear zone. The width of the shear zone as defined by the stress field is about 15 grain diameters.

ops as finite strain accumulates across the shear zone, and is therefore only a feature of mature shear zones. Similar behaviour has been shown in numerical simulations using continuum mechanics²⁴ with an elastic-plastic rheology.

To attempt a comparison between the simulations and nature, it is necessary to identify the spatial scale represented by the model system. A definitive statement cannot be made because the simulations have a narrow range of grain sizes, whereas 'grain sizes' in nature have a broad, self-similar distribution²⁵ ranging from individual minerals up to crustal blocks. For the present, consider a hypothetical system containing a wide range of grain

sizes. It is suggested that the width of a shear zone showing stress rotation (Fig. 4b) would be controlled by the larger grain sizes in the distribution, because it is necessary to mobilize grains of all sizes to accommodate the build-up of inelastic strain in a continuous zone. It is also suggested that the width of the shear dislocation for an individual earthquake (Fig. 3b) would be controlled by the smaller grain sizes in the distribution. A comparison of Figs 3b and 4b provides support for these suggestions but is not conclusive.

If correct, the model proposed here removes the need for two relatively complex ideas that have been widely considered. The

first is that large faults have anomalously low frictional resistance to sliding, most probably due to high pore-fluid pressures confined to the fault zone⁶; this is proposed as a solution to the San Andreas stress/heat-flow paradox²⁶. A difficulty with the weak fault model is that the smaller faults around a major fault appear to be strong, so that the fluid pressure must only be elevated within the major fault. It may be argued that the required permeability structure is contrived, particularly in a region of active strike-slip tectonics. The model proposed here involves no variation in pore fluid pressure, produces no frictional heating (by definition), and demonstrates how stress trajectories can be oriented towards the perpendicular as they cross a mature shear zone.

The second idea is that rate-dependent friction²⁷ or slip-dependent strength²⁸ of faults is needed for earthquake nucleation. In the DEM simulations, the heterogeneity of the material allows for local dilation although the overall state of stress is compressive. The build-up of elastic strain amplifies this heterogeneity and at a critical level produces local grain separation and slip. Elastodynamic interaction then promotes further amplification of heterogeneity and produces the observed macroscopic failure. This concept is consistent with recent seismological²⁹, experimental³⁰ and numerical¹³ studies, all of which support the idea that earthquake rupture propagates across a fault as a dilatant slip pulse. In a static sense, the failure process in these simulations involves the geometrical collapse of load-bearing grain bridges at a critical strain³¹.

Seismological estimates indicate that only a small fraction of the strain energy from an earthquake is radiated⁵. A prediction of this model, and any model with little or no frictional fault heating, is that most of the strain energy from earthquakes is released as elastic waves. It may be possible to resolve this contradiction as follows. In the DEM simulations, the dominant wavelength in the energy radiated by each AE is the same as the scattering length in the surrounding heterogeneous material. Only a fraction of the energy is carried by coherent, longer-wavelength radiation; the remainder is trapped around the source by multiple scattering and is gradually dissipated by viscous damping. To speculate that a similar process is occurring in nature implies that the majority of high-frequency seismic energy escapes detection, perhaps due to severe attenuation concentrated in the shallow crust³². Identifying this energy, by acquiring high-frequency, near-field seismic data, may provide a test of the model. Dissipation of this energy might also create the broad (100-km-wide) heat flow anomaly observed across the San Andreas fault³ (as distinct from the 10-km-wide anomaly expected from frictional fault heating). There are parallels here with the ideas of 'acoustic fluidisation' in fault zones during earthquakes³³, Anderson localization in granular materials³⁴, and the barrier model of heterogeneous seismic sources³⁵, all of which involve elastic wave propagation in heterogeneous materials. □

Received 11 September 1995; accepted 8 May 1996.

- Byerlee, J. *Pure appl. Geophys.* **116**, 615–626 (1976).
- Lachenbruch, A. H. & Sass, J. H. *J. geophys. Res.* **85**, 6185–6222 (1980).
- Lachenbruch, A. H. & Sass, J. H. *J. geophys. Res.* **97**, 4995–5015 (1992).
- Zoback, M. D. et al. *Science* **238**, 1105–1111 (1987).
- Kanamori, H. A. *Rev. Earth planet. Sci.* **22**, 207–237 (1994).
- Rice, J. R. in *Fault Mechanics and Transport Properties in Rocks* (eds Evans, B. & Wong T.-F.) 475–503 (Academic, San Diego, 1993).
- King, G. C. P. *Pure appl. Geophys.* **121**, 761–814 (1983).
- Cundall, P. A. & Strack, O. D. L. *Geotechnique* **29**, 47–65 (1979).
- Corkum, B. T. & Ting, J. M. *Publication 86–11* (Dept of Civil Engng, Univ. Toronto, 1986).
- Poschel, T. & Buchholtz, V. *Phys. Rev. Lett.* **71**, 3963–3966 (1993).
- Lockner, D. *Int. J. Rock* **30**, 883–899 (1993).
- Scott, D. R. *Eos* (abstr.) **75**, 614 (1994).
- Mora, P. & Place, D. *Pure appl. Geophys.* **143**, 61–87 (1994).
- Yoshida, T., Tatsuoka, F., Siddiquee, M. S. A., Kamegai, Y. & Park, C. S. in *Localisation and Bifurcation Theory for Soils and Rocks* (eds Chambon, R., Desrué, J. & Vardoulakis, I.) 165–179 (A. A. Balkema, Rotterdam, 1994).
- Ord, A., Vardoulakis, I. & Kajewski, R. *Int. J. Rock* **28**, 397–409 (1991).
- Cundall, P. A. *Ing. Arch.* **59**, 148–159 (1989).
- Bardet, J.-P. & Proubet, J.-P. *Geotechnique* **41**, 599–613 (1991).
- Hobbs, B. E. & Ord, A. *Ing. Arch.* **59**, 209–220 (1989).
- Poljakov, A. N. B., Hermann, H. J., Podladnikov, Y. Y. & Roux, S. *Fractals* **2**, 567–581 (1994).
- Rudnicki, J. W. & Rice, J. R. *Int. J. Mech. Phys. Solids* **23**, 371–394 (1975).
- Vardoulakis, I. *Ing. Arch.* **59**, 106–113 (1989).
- Mühlhaus, H. B. & Vardoulakis, I. *Geotechnique* **37**, 271–283 (1987).

- Massonnet, D. et al. *Nature* **364**, 138–142 (1993).
- Cundall, P. A. in *Mechanics of Jointed and Faulted Rock* (ed. Rossmanith, H.-P.) 11–18 (Balkema, Rotterdam, 1990).
- Stearcy, S. J. & Sammis, C. G. *Nature* **353**, 250–252 (1991).
- Hickman, S. H. *Rev. Geophys. Suppl. Ser.* **29**, 759–775 (1991).
- Dieterich, J. H. *Tectonophysics* **211**, 115–134 (1992).
- Ohnaka, M. *Tectonophysics* **211**, 149–178 (1992).
- Heaton, T. H. *Phys. Earth planet. Inter.* **64**, 1–20 (1990).
- Brune, J. N., Brown, S. & Johnson, P. A. *Tectonophysics* **218**, 59–67 (1993).
- Sammis, C. G. & Steacy, S. J. *Pure appl. Geophys.* **142**, 777–794 (1994).
- Abercrombie, R. & Leary, P. *Geophys. Res. Lett.* **20**, 1511–1514 (1993).
- Melosh, H. J. *Nature* **379**, 601–606 (1996).
- Sornette, A., Sornette, D. & Evesque, P. *Nonlin. Process. Geophys.* **1**, 209–218 (1994).
- Aki, K., Bouchon, M., Chouet, B. & Das, S. *Ann. Geofis.* **XX**, 341 (1977).
- Liu, C.-h. et al. *Science* **269**, 513–515 (1995).

ACKNOWLEDGEMENTS. I thank C. G. Sammis for sustained encouragement and support. This work was supported in part by an NERC Research Fellowship, an NERC small grant and an NSF research grant.

Detection of tidal dissipation in the solid Earth by satellite tracking and altimetry

R. D. Ray*, R. J. Eanes† & B. F. Chao‡

* HSTX, NASA Goddard Space Flight Center, Code 926, Greenbelt, Maryland 20771, USA

† Center for Space Research, University of Texas, Austin, Texas 78712, USA

‡ NASA Goddard Space Flight Center, Code 926, Greenbelt, Maryland 20771, USA

THE rate at which tidal energy is dissipated in the solid Earth can constrain the anelastic properties of the Earth at frequencies much lower than those accessible with seismology. The dissipative properties of a system are usually expressed as a 'quality factor', Q ; estimates^{1–7} of the semi-diurnal solid-Earth Q range from 90 to 500. But observational constraints on this quantity are difficult to obtain, because dissipation by the body tide is masked by the much greater dissipation that occurs in the oceans^{8,9}. Here we show that recent accurate measurements of the ocean tide obtained by the Topex/Poseidon satellite altimeter¹⁰, combined with nearly two decades of laser tracking of satellite orbit perturbations¹¹ (which are sensitive to the total planetary dissipation rate), permit an estimate of the solid-Earth dissipation rate at semi-diurnal period. We find that the body tide lags the principal lunar tidal potential by $0.16 \pm 0.09^\circ$, implying a dissipation rate of 83 ± 45 gigawatts and a solid-Earth Q of 370 at a period of 12.4 hours. The observed lag agrees well with Zschau's 'most probable' lag¹ of 0.21° (deduced from observations of the Chandler wobble), and favours the higher values of Q estimated theoretically.

The Moon and Sun raise tides in the oceans, the atmosphere and the solid Earth. All three, but primarily the oceans, dissipate some amount of the imparted tidal energy; an accurate accounting of the tidal energy balance has been a long-sought goal¹². The problem is of some importance, not only because the total dissipation rate is so large—approaching 4×10^{12} watts (ref. 9)—but also because the problem bears on the physics of the dissipation mechanisms, and hence on the dynamics of the oceans and atmosphere and on the anelastic structure of the Earth.

The most elegant, and so far most accurate, method for determining the planetary (combined ocean-atmosphere-Earth) tidal dissipation rate is to analyse the tidally induced perturbations in the orbits of artificial satellites: several 'geodetic' satellites, ideal for this task, have now been routinely tracked by lasers for two decades. From the planetary rate, the solid-Earth dissipation is found by subtracting the oceanic and atmospheric components, using altimetric observations of the ocean and barometric (surface) observations of the atmosphere. Complications arise primarily from details of the tracking and altimeter

analyses, and from the fact that the solid-Earth component of dissipation is a very small residual signal, necessitating highly accurate and precise tidal observations. For such reasons, we concentrate here exclusively on the principal lunar semidiurnal tide M_2 , which is the tide most accurately determined. The solar S_2 tide is confounded by non-gravitational (radiational and thermodynamic) effects, whereas the smaller diurnal tides are, at present, less well determined.

As the Earth, ocean and atmosphere respond to the Moon's primary tide-generating potential, Φ^P , the induced deformations result in a secondary potential Φ^S , to which the satellite orbit perturbations are directly sensitive. Dissipation in the planet is manifested by a lag in the Earth's response, and hence by a lag of Φ^S relative to Φ^P . Thus, as shown by Platzman¹³, the two potentials, combined in a surface integral over the globe, determine the rate of tidal working, and hence the planetary dissipation:

$$W = \frac{5}{4\pi Ga} \left\langle \Phi^P \frac{\partial \Phi^S}{\partial t} \right\rangle dS \quad (1)$$

where G is the gravitational constant, a the Earth's mean radius, and $\langle \rangle$ brackets indicate averaging over one tidal cycle. Because Φ^P for M_2 is a spherical harmonic function of degree and order 2, orthogonality properties of equation (1) allow us to restrict attention to the same degree-2 components of the tidal deformations. (A very small component of the solid-Earth dissipation is

dependent on all surface harmonics of the fluid tides: Platzman's $R(h' - k')$ integral, which we neglect.) Here, we adopt tidal coefficients from five recent satellite tracking and altimetric analyses. These coefficients, listed in Table 1, give the amplitude D_{22} and phase lag ψ_{22} of the degree-2, order-2 (nominal) ocean tidal height. The tracking-based coefficients were determined as part of large least-squares inversions for both static and time-varying components of the Earth's gravity field, using many years of tracking observations. The altimetric coefficients were determined by numerical quadrature of global tidal elevations fields that had been directly measured by the Topex/Poseidon satellite altimeter.

In the published literature, as well as in our Table 1, the quoted coefficients correspond to a 'nominal' ocean tide, because they are based on elastic Earth models and a neglected atmosphere. In fact, both the atmospheric tide and anelasticity in the Earth affects the satellite-tracking coefficients (and to far less an extent, anelasticity affects the altimetric coefficients). It is primarily the anelasticity that accounts for the offset between the tracking and altimeter coefficients evident in the ordinate of Fig. 1 (right panel), and it is this systematic discrepancy that allows us, for the first time, to set credible observational constraints on the body-tide lag, and hence the dissipation.

As is evident from Fig. 1 (left panel), ocean-tide models developed before Topex/Poseidon are inadequate for our task.

Numerical hydrodynamic tide models have always been plagued by inaccuracies arising from inadequate bathymetric data, unknown frictional and viscosity parameters, or physical modelling that is too simplistic. Direct tidal measurements by coastal and bottom gauges are valuable but too sparse to attain required global accuracies, whereas pre-Topex satellite missions¹⁴ have been limited by numerous systematic errors. Topex/Poseidon overcomes most of these difficulties: its orbit was purposely configured to minimize tidal aliasing problems¹⁵ and its tracking systems have yielded orbital ephemerides accurate to about 3 cm r.m.s. (ref. 16). Tidal studies resulting from Topex/Poseidon have been reviewed in refs 17,18.

Our three altimetric tide models, although based on essentially the same data, use different processing and analysis methods. Two models use binning methods¹⁴ which hamper meaningful error estimates, but the model of Ray *et al.* ref. 25 is based on a single global inversion for the coefficients in a normal-mode expansion. The resulting global error covariance matrix has been 'calibrated' (that is, reweighted for unmodelled systematic or correlated errors by generating and comparing subset solutions¹⁹) and then propagated into estimated errors for the (2,2) coefficients. The tabulated 0.036 cm is therefore a realistic uncertainty. Some further corroboration of this can be had from the altimeter-based frequency-wavenumber spectra of ocean variability recently computed by Wunsch and Stammer²⁰. Their diagrams (their Fig. 2a) display a tidal 'ridge' caused by tide-model errors and, using the most generous of assumptions, we find that the size of this ridge suggests an error in the (CSR3.0) degree-2 coefficient of order 0.02 cm.

The following analysis is based loosely on the earlier work of Zschau¹ and Lambeck³. Let ε be a small lag angle expressing the phase delay between the body tide and the tidal potential Φ^P ; the goal is to determine ε . The quadrature part of Φ^S (corresponding to the vertical axis in Fig. 1) is determined from the tracking coefficient $D_{22}^T \sin \psi_{22}^T$. Φ^S is then decomposed into three parts: $\Phi^S = \Phi_B^S + \Phi_{O+L}^S + \Phi_{A-L}^S$ arising, respectively, from the body tide, the ocean tide and its corresponding load displacement, and the air tide and its load. Writing out these relationships, one

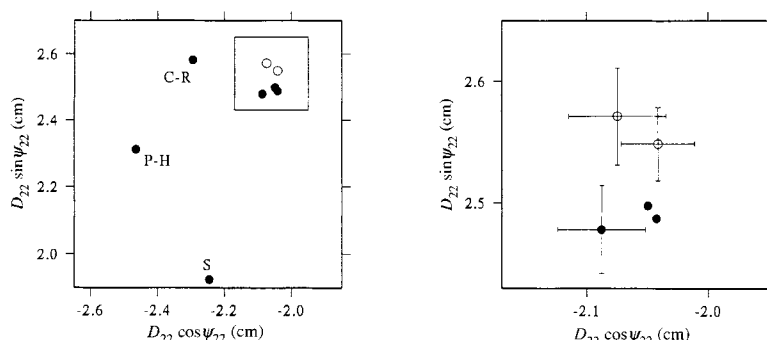


FIG. 1 In-phase and quadrature components of degree-2, order-2 prograde M_2 coefficients of the 'nominal' ocean tide. Left, three older tide models from S: Schwiderski³¹; P-H: Parke and Hendershott³²; C-R: Cartwright and Ray¹⁴. Right, zoom view for the estimates given in Table 1. Satellite-tracking coefficients are unfilled circles; others are either numerical hydrodynamic or altimetric. 1σ standard errors are plotted where available.

TABLE 1 Degree-2 order-2 prograde harmonics of the M_2 ocean tide

	Version	D_{22} (cm)	ψ_{22} (deg)	σ (cm)
Topex/Poseidon analyses				
Ray <i>et al.</i> ²⁵	OP950310b	3.241	130.11°	0.036
Eanes ²⁶	CSR3.0	3.218	129.40°	—
Schrama and Ray ¹⁵	950308	3.231	129.38°	—
Satellite-tracking analyses				
Lerch <i>et al.</i> ²⁷	GEM-T3S	3.30	128.9°	0.04
Eanes	LLA95	3.265	128.70°	0.03

These coefficients describe the spherical harmonic component of the nominal ocean tide height: $D_{22} P_2^2(\sin \varphi) \cos(\omega t + 2\lambda - \psi_{22})$, where (φ, λ) are latitude and longitude, and ω is the M_2 tidal frequency. Tracking solution GEM-T3S is from an analysis of laser, Doppler, and optical tracking of 31 different satellites; its error estimate has been carefully calibrated¹⁹. Solution LLA95 is from a 19-year span of Lageos-1 laser data and is an updated version of ref. 28. All tracking solutions have been adjusted to conform to a consistent set of constants: $\rho_w = 1.035 \text{ kg m}^{-3}$, $a = 6.371 \text{ km}$ and $k_2' = -0.309$ (refs 24,29). The three altimetric models have been supplemented by various numerical hydrodynamic models in the polar and marginal seas, so the above coefficients correspond to truly global oceans, not limited by the Topex/Poseidon sampling. The model of Le Provost *et al.*³⁰ was used for all latitudes above 66°. The supplemental models, although undoubtedly of benefit, affect the harmonic coefficients by well less than the quoted 0.036-cm uncertainty.

arrives at the following equation for ε :

$$C_1 D_{22}^T \sin \psi_{22}^T = k_2 \sin \varepsilon + C_1 D_{22}^A \sin (\psi_{22}^A + \delta) \\ - C_2 \sin \varepsilon + C_3 E_{22}^{\text{air}} \sin (\gamma_{22}^{\text{air}} + \delta) \quad (2)$$

where $C_1 = 3\rho_w(1+k'_2)/(5\rho_e\bar{H}) = 0.9590\text{ m}^{-1}$, $C_2 = 3\rho_w h_2(1+k'_2)/(5\rho_e) = 4.752 \times 10^{-2}$, and $C_3 = 3(1+k'_2)/(5g\rho_e\bar{H}) = 9.445 \times 10^{-5}\text{ Pa}^{-1}$; k_2 and k'_2 are the potential Love and loading numbers (constants that describe how Φ^S depends on the Earth tide and on Earth loading¹²), h_2 the displacement Love number (describing the vertical displacement of the Earth tide¹²), ρ_w and ρ_e are the densities of sea water and Earth, g the gravitational acceleration, and $\bar{H}$ for M_2 equals 8.137 cm. D_{22}^T , ψ_{22}^T are the coefficients of the altimeter-derived ocean tide (which were derived assuming that $\varepsilon = 0$ and therefore required the small C_2 correction term), and E_{22}^{air} , γ_{22}^{air} are the coefficients of the lunar barometric tide^{21,22}. Angle δ represents an additional phase offset in the ocean plus load tide and the air plus load tide, induced by anelastic response to Earth loading. Because equation (2) is fairly insensitive to δ , we hereafter set it to zero. (If δ were as large as 1° , its neglect would affect the estimated ε by only 0.07° . Further, there is no physical reason to expect that $\delta \gg \varepsilon$; theoretical calculations based on various Earth models^{23,24} suggest that $\delta < 0.02^\circ$.) It is straightforward to show that ε is also insensitive to the other adopted constants required in C_1 , C_2 and C_3 .

Equation (2) is analogous to the simpler equation (8) of Zschau¹. Zschau was able to write an additional equation for the in-phase component ($k_2 \cos \varepsilon$) and thereby estimated both k_2 and ε (at least in principle; in practice the ocean models available to him were too inaccurate for this). Here, because of our use of altimetry, we cannot estimate k_2 independently of h_2 . Alternatively, we may use the in-phase components to place bounds on δ . This must be examined in more detail elsewhere, but the bounds turn out to be weak and, given the lack of any systematic in-phase difference in Fig. 1, the result is not significantly different from zero.

For the tidal coefficients D_{22}^T , ψ_{22}^T , D_{22}^A , ψ_{22}^A , we adopt the means of the appropriate models listed in Table 1. For the tracking and altimeter standard errors, we use 0.024 and 0.036 cm, respectively; the latter is simply the listed value for the one model and ignores averaging as all three altimeter models use essentially the same data. Solution of equation (2) then yields $\varepsilon = 0.16^\circ \pm 0.09^\circ$. This value of ε implies¹³ that friction in the Earth's body tide consumes 83 ± 45 gigawatts of tidal power. For comparison, this is $\sim 3\%$ of the ocean's dissipation, and, according to Platzman's estimate²², about eight times the atmosphere's dissipation. If Q^{-1} is the body tide's specific dissipation function, then the effective tidal $Q = 1/\tan \varepsilon$ is found to be 370, with approximate confidence limits of (200,800).

This estimate of Q helps fill an important observational gap between seismic modes of period less than 1 hour and the Chandler wobble of period 14 months; even measurements with relatively wide confidence bounds may help clarify previous disparate theoretical and/or indirect estimates. For example, one indirect approach² infers tidal Q from free-oscillation data, arguing that the body tide and the ${}_0S_2$ free-oscillation mode should, aside from a slight frequency dependence, display similar Q (although Bostrom³ suggests that the Earth's rotation may reduce the tidal Q). Munk⁴, in fact, adopted a free-oscillation Q of 350 (now thought to be too low⁵) to estimate a solid-Earth tidal dissipation essentially equivalent to our observational estimate. Other theoretical estimates of complex k_2 Love numbers⁶ suggest bounds on tidal Q of (90, 500), with a preference around 210, whereas another⁷ k_2 value suggests a Q of 126, which is outside our (1σ) observational range. Finally, Zschau¹, adopting a single absorption-band model between seismic and Chandler wobble periods, arrives at 'most probable' estimates for ε of 0.21° and solid-Earth dissipation of 120 GW. Within their error bounds, our observations are consistent with Zschau's values and with the assumed smooth frequency dependence between seismic and Chandler periods. $\square$

Received 19 January; accepted 8 May 1996.

1. Zschau, J. in *Space Geodesy and Geodynamics* (eds Anderson, A. & Cazenave, A.) 315–344 (Academic, London, 1986).
2. Lagus, P. L. & Anderson, D. L. *Phys. Earth planet. Inter.* **1**, 505–510 (1968).
3. Bostrom, R. C. *EOS* **76**, F60 (1995).
4. Munk, W. Q. *Jl R. astr. Soc.* **9**, 352–375 (1968).
5. Tanimoto, T. *Geophys. Res. Lett.* **17**, 669–672 (1990).
6. Wahr, J. & Bergen, Z. *Geophys. J. R. astr. Soc.* **87**, 633–668 (1986).
7. Dehant, V. *Phys. Earth planet. Inter.* **49**, 97–116 (1987).
8. Lambeck, K. *Phil. Trans. R. Soc. Lond.* **A287**, 545–594 (1977).
9. Ray, R. D. in *The Oceans* (eds Majumdar, S. K. et al.) 171–185 (Pennsylvania Academy of Science, Easton, 1994).
10. Fu, L.-L. et al. *J. geophys. Res.* **99**, 24369–24381 (1994).
11. Cohen, S. C. & Smith, D. E. *J. geophys. Res.* **90**, 9217–9220 (1985).
12. Munk, W. H. & MacDonald, G. J. F. *The Rotation of the Earth* (Cambridge Univ. Press, 1960).
13. Platzman, G. W. *Rev. Geophys. Space Phys.* **22**, 73–84 (1984).
14. Cartwright, D. E. & Ray, R. D. *J. geophys. Res.* **96**, 16897–16912 (1991).
15. Schrama, E. J. O. & Ray, R. D. *J. geophys. Res.* **99**, 24799–24808 (1994).
16. Marshall, J. A. et al. *J. geophys. Res.* **100**, 25331–25352 (1995).
17. Le Provost, C., Bennett, A. F. & Cartwright, D. E. *Science* **267**, 639–642 (1995).
18. Andersen, O. B., Woodworth, P. L. & Flather, R. A. *J. geophys. Res.* **100**, 25261–25282 (1995).
19. Lerch, F. *Bull. Géod.* **65**, 44–52 (1991).
20. Wunsch, C. & Stammer, D. *J. geophys. Res.* **100**, 24895–24910 (1995).
21. Haurwitz, B. & Cowley, A. D. *Pure appl. Geophys.* **102**, 193–222 (1973).
22. Platzman, G. W. *Pure Appl. Geophys.* **137**, 1–33 (1991).
23. Zschau, J. in *Tidal Friction and the Earth's Rotation* (eds Brosche, P. & Sündermann, J.) 62–94 (Springer, New York, 1978).
24. Pagiatas, S. D. *Geophys. J. Int.* **103**, 541–560 (1990).
25. Ray, R. D., Sanchez, B. & Cartwright, D. E. *EOS* **75**, 108 (1994).
26. Eanes, R. J. *EOS* **75**, 108 (1994).
27. Lerch, F. J. et al. NASA Tech. Memo. 104555 (Goddard Space Flight Center, Greenbelt, 1992).
28. Tapley, B. D., Schutz, B. E., Eanes, R. J., Ries, J. C. & Watkins, M. M. in *Contributions of Space Geodesy to Geodynamics: Earth Dynamics* (eds Smith, D. E. & Turcotte, D.) 147–174 (Am. geophys. Union, Washington, 1993).
29. Gill, A. *Atmosphere-Ocean Dynamics* (Academic, New York, 1982).
30. Le Provost, C., Genco, M. L., Lyard, F., Vincent, P. & Canceil, P. *J. geophys. Res.* **99**, 24777–24797 (1994).
31. Schwiderski, E. W. *Mar. Geol.* **6**, 219–265 (1983).
32. Parke, M. E. & Hendershott, M. C. *Mar. Geol.* **3**, 379–407 (1980).

ACKNOWLEDGEMENTS. We thank C. F. Yoder and especially G. W. Platzman for useful discussions, and C. Wunsch for his frequency-wavenumber spectra.

CORRESPONDENCE should be addressed to R.D.R. (e-mail: ray@nemo.gsfc.nasa.gov).

The guinea-pig is not a rodent

Anna Maria D'Erchia*†, Carmela Gissi*†, Graziano Pesole‡, Cecilia Saccone*§ & Ulfur Arnason†

* Dipartimento di Biochimica e Biologia Molecolare, Università di Bari, 70125 Bari, Italy

† Department of Evolutionary Molecular Systematics, University of Lund, Sölvegatan 29, S-22362 Lund, Sweden

‡ Dipartimento di Biologia DBAF, Università della Basilicata, 30100 Potenza, Italy

§ Centro di Studio sui Mitochondri e Metabolismo Energetico, Consiglio Nazionale delle Ricerche, 70125 Bari, Italy

In 1991 Graur *et al.* raised the question of whether the guinea-pig, *Cavia porcellus*, is a rodent¹. They suggested that the guinea-pig and myomorph rodents diverged before the separation between myomorph rodents and a lineage leading to primates and artiodactyls. Several findings have since been reported, both for and against this phylogeny, thereby highlighting the issue of the validity of molecular analysis in mammalian phylogeny. Here we present findings based on the sequence of the complete mitochondrial genome of the guinea-pig, which strongly contradict rodent monophyly. The conclusions are based on the cumulative evidence provided by orthologically inherited genes and the use of three different analytical methods, none of which joins the guinea-pig with myomorph rodents. In addition to the phylogenetic conclusions, we also draw attention to several factors that are important for the validity of phylogenetic analysis based on molecular data.

The order Rodentia is by far the most speciose mammalian order, comprising 1,814 species and 29 families. It is traditionally divided into three suborders: Sciuromorpha (squirrels), Myomorpha (mice, rats), and Hysticomorpha (porcupines, guinea-pigs).

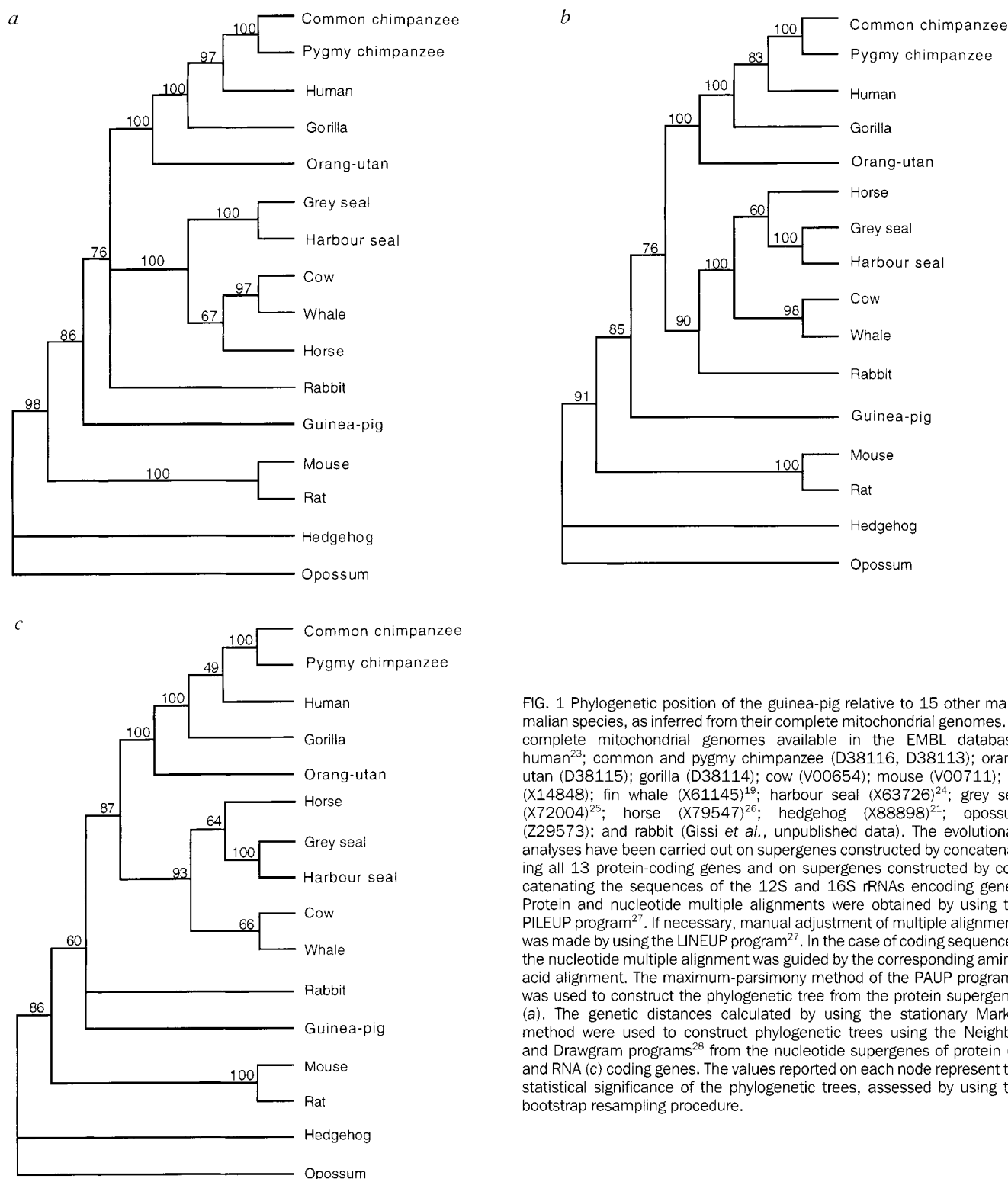


FIG. 1 Phylogenetic position of the guinea-pig relative to 15 other mammalian species, as inferred from their complete mitochondrial genomes. All complete mitochondrial genomes available in the EMBL database: human²³; common and pygmy chimpanzee (D38116, D38113); orang-utan (D38115); gorilla (D38114); cow (V00654); mouse (V00711); rat (X14848); fin whale (X61145)¹⁹; harbour seal (X63726)²⁴; grey seal (X72004)²⁵; horse (X79547)²⁶; hedgehog (X88898)²¹; opossum (Z29573); and rabbit (Gissi *et al.*, unpublished data). The evolutionary analyses have been carried out on supergenes constructed by concatenating all 13 protein-coding genes and on supergenes constructed by concatenating the sequences of the 12S and 16S rRNAs encoding genes. Protein and nucleotide multiple alignments were obtained by using the PILEUP program²⁷. If necessary, manual adjustment of multiple alignments was made by using the LINEUP program²⁷. In the case of coding sequences, the nucleotide multiple alignment was guided by the corresponding amino-acid alignment. The maximum-parsimony method of the PAUP program¹⁵ was used to construct the phylogenetic tree from the protein supergenes (a). The genetic distances calculated by using the stationary Markov method were used to construct phylogenetic trees using the Neighbor and Drawgram programs²⁸ from the nucleotide supergenes of protein (b) and RNA (c) coding genes. The values reported on each node represent the statistical significance of the phylogenetic trees, assessed by using the bootstrap resampling procedure.

The guinea-pig, *Cavia porcellus*, is generally classified as a New World hystricomorph rodent belonging to the family Caviomorpha. The joining of the three rodent suborders into a single clade was supported by morphological data. This view had not been questioned until Graur *et al.*¹ concluded, on the basis of maximum parsimony analysis of 15 protein sequences, that the order Rodentia is polyphyletic, and suggested that the suborder Hystricomorpha should be separated from the Rodentia and given the status of an order, Hystricomorpha. The phylogenetic position of the guinea-pig relative to other rodents, and hence also rodent phylogeny in general, has since become a matter of considerable

contention among evolutionists, and rodent monophyly has been questioned by analyses of other sets of nucleotide and protein sequences²⁻⁶. However, some of the former set of sequence data have been re-examined using the maximum-likelihood method^{7,8}, and 19 protein sequences using four groups of taxa⁹, and sequences of mitochondrial 12S and 16S ribosomal RNA genes¹⁰ have been analysed, all with inconclusive results. A different approach, based on studies of BC1 RNA expression, a retroposon element that is present in all rodents and guinea-pig but not in some other mammals, has again suggested rodent monophyly¹¹.

These inconsistent, and even contradictory, findings show that the evolutionary position of the guinea-pig has not been unequivocally clarified, although it cannot be excluded that some of the controversies may be due to the different analytical approaches used, insufficient number of taxa examined, or to the inadequate length of the sequences used in the phylogenetic analyses.

In an attempt to shed light on this problem, we have sequenced the complete mitochondrial genome (mtDNA) of the guinea-pig. We chose mtDNA because of its maternal inheritance, lack of recombination, and especially the presence of orthologous genes that make it a particularly suitable and powerful tool for phylogenetic analysis.

The study has included 14 complete mammalian mtDNA sequences available in the literature, representing eight acknowledged orders: Primates, Artiodactyla, Cetacea, Perissodactyla, Carnivora, Rodentia, Insectivora and Marsupialia (outgroup). We have also included mtDNA from the rabbit (Lagomorpha), the sequence of which was recently determined (C.G. *et al.*, unpublished results). On morphological grounds, the Lagomorpha is claimed to form a monophyletic clade with Rodentia (the so-called Glires)¹², but molecular data do not support this relationship^{13,14}. The inclusion of Lagomorpha is therefore particularly relevant in this evolutionary analysis.

To minimize the methodological bias, we performed the phylogenetic analyses by using three completely different methods: (1) the maximum-parsimony method¹⁵, a deterministic method that we applied to the derived protein sequences; (2) the protein maximum-likelihood method¹⁶; and (3) the stationary Markov model¹⁷, a stochastic method that we applied to nucleotide sequences.

The phylogenetic tree (maximum parsimony analysis) based on the concatenated sequences of all 13 protein-coding mitochondrial genes is shown in Fig. 1a. A clade comprising Lagomorpha, Primates, Carnivora, Perissodactyla, Artiodactyla and Cetacea is supported by a bootstrap value of 76%. Consistent with previous findings there is a distinct affinity between Artiodactyla and Cetacea^{18–21}. The position of Lagomorpha remains unsettled, as the low bootstrap value does not allow us to resolve the trichotomy among Lagomorpha, Primates and the clade including Carnivora, Perissodactyla, Artiodactyla and Cetacea. The guinea-pig falls outside the myomorph rodents (mouse, rat) as a sister-group of the clade including Lagomorpha, Primates, Carnivora, Perissodactyla, Artiodactyla and Cetacea. The alternative topology with a monophyletic Rodentia has a bootstrap support value of only 1%. Insectivora, represented by the hedgehog, *Erinaceus europaeus*, appears at the basal branch of the eutherian tree, consistent with previous results²¹. The analysis of protein supergenes using the maximum-likelihood method¹⁶ results essentially in the same tree topology (data not shown), with Lagomorpha more closely related to the clade including Artiodactyla, Cetacea, Perissodactyla and Carnivora.

The phylogenetic tree obtained by applying the neighbour-joining method²² on the pairwise distances of the concatenated nucleotide sequences of the 13 protein-coding genes, as calculated by applying the stationary Markov model on synonymous nucleotide substitutions, is shown in Fig. 1b. Synonymous substitutions have not been included in the analysis because they show a marked compositional heterogeneity among mammals (lack of stationarity), and also, because of their fast substitution rate, the nucleotide divergence is well above the saturation level for many interspecies comparisons. The close relationship between the guinea-pig and the Lagomorpha/Primates/Carnivora/Perissodactyla/Artiodactyla(+Cetacea) clade is again supported by high bootstrap values (85%) (Fig. 1b), whereas the alternative tree with a monophyletic Rodentia has a bootstrap value of 10%.

The phylogenetic tree obtained by using the stationary Markov model/neighbour-joining method on the rRNA supergenes is shown in Fig. 1c. It reproduces results reported in Fig 1a, b, further supporting the phylogenetic relationships described above.

Our analyses of the phylogenetic position of the guinea-pig were performed using a comprehensive data set, both with respect to the number of orthologous genes and to the number of mammalian orders examined. The different analytical approaches and sets of data yielded congruent results that were consistent with the hypothesis of Rodent polyphyly¹. It should be noted, however, that our trees, which identify the guinea-pig as an outgroup to Lagomorpha/Primates/Carnivora/Perissodactyla/Artiodactyla(+Cetacea), differ slightly from that proposed by Graur *et al.*, in which the guinea pig constitute an outgroup to myomorph rodents, primates and artiodactyls¹.

Phylogenetic relationships among Rodentia (*sensu lato*) have been addressed by analysing nuclear⁹ and mitochondrial rRNA genes¹⁰. Regarding the analysis of nuclear genes⁹, it is worth noting that, although the findings, in general, do not support rodent polyphyly, different genes have provided dissimilar answers to the question of rodent phylogeny. Such discrepancies are often observed when using nuclear genes, perhaps because some genes evolve under different evolutionary constraints in the various tree branches, or because the genes analysed might be paralogous. In contrast, mitochondrial genomes contain only orthologous single-copy genes, and can thus provide more reliable phylogenies.

Our results obtained with mt rRNA (Fig. 1c), which confirm the findings based on protein-coding genes, are inconsistent with previous findings¹⁰ that support rodent monophyly by analysing these same genes. These differences might be due to differences in the multiple alignment, which in the case of rRNA genes is not obvious, rather than to the analytical approach applied.

The close relationship we found between Cetacea and Artiodactyla is strongly supported, confirming previous molecular data^{18–21}. The positions of Perissodactyla and Lagomorpha remain uncertain, but a clade including Cetacea, Artiodactyla, Carnivora and Perissodactyla is strongly supported. Our data contradict morphological studies¹² that place Rodentia and Lagomorpha in the monophyletic cohort Glires, but agree with other molecular evolutionary analyses based on nuclear encoded proteins¹⁴ that place Lagomorpha closer to Primates than to Rodentia; both analyses reject the monophyletic Glires. The stationary Markov model/neighbour-joining and the protein maximum-likelihood findings, which place Lagomorpha closely akin to the clade including Artiodactyla, Cetacea, Perissodactyla and Carnivora, are highly suggestive, but more data are necessary before this relation can be considered reliable.

The phylogenetic position of the guinea-pig is highly reliable, as different methodological approaches provide highly consistent results. The large number of sites included, and the very significant bootstrap values obtained, provide strong evidence in favour of the inclusion of guinea-pig in a new mammalian order distinct from Rodentia.

Because of the exceptional complexity of the order Rodentia, sequence data are still limited, and the phylogeny will be clearer when a larger number of species can be considered in phylogenetic analyses. □

Received 9 November 1995; accepted 8 May 1996.

1. Graur, D., Hide, W. A. & Li, W. H. *Nature* **351**, 649–652 (1991).
2. Li, W. H., Hide, W. A., Zharkikh, A., Ma, D. P. & Graur, D. *J. Heredity* **83**, 174–181 (1992).
3. Graur, D., Hide, W. A., Zharkikh, A. & Li, W. H. *Comp. Biochem. Physiol.* **101B**, 495–498 (1992).
4. Wolf, B., Reinecke, K., Aumann, K. D., Brigelius-Flohé, R. & Flohé, L. *Biol. Chem. Hoppe-Seyler* **374**, 641–649 (1993).
5. Noguchi, T., Fujiwara, S., Hayashi, S. & Sakuraba, H. *Comp. Biochem. Physiol.* **107**, 179–182 (1994).
6. Ma, D. P., Zharkikh, A., Graur, D., VandeBerg, J. L. & Li, W. H. *J. molec. Evol.* **36**, 327–334 (1993).
7. Hasegawa, M., Cao, Y., Adachi, J. & Yano, T. *Nature* **355**, 595 (1992).
8. Cao, Y., Adachi, J., Yano, T. & Hasegawa, M. *Molec. Biol. Evol.* **11**, 593–604 (1994).
9. Kuma, K. & Miyata, T. *Jap. J. Genet.* **69**, 555–566 (1994).
10. Frye, M. S. & Hedges, S. B. *Molec. Biol. Evol.* **12**(1), 168–176 (1995).
11. Martignetti, J. A. & Brosius, J. *Proc. natn. Acad. Sci. U.S.A.* **90**, 9698–9702 (1993).
12. Novacek, N. J. *Nature* **356**, 121–125 (1992).
13. Graur, D. *FEBS Lett.* **325**, 152–159 (1993).
14. Graur, D., Duret, L. & Gouy, M. *Nature* **379**, 333–335 (1996).
15. Swofford, D. L. *PAUP: Phylogenetic Analysis Using Parsimony Version 3.1.1* (Illinois Natural History Survey, Champaign, 1993).

16. Adachi, J. & Hasegawa, M. *MOLPHY: Programs for Molecular Phylogenetics 2.2* (Computer Science Monographs, No. 27, Institute of Statistical Mathematics, Tokyo, 1992).
17. Saccone, C., Lanave, C., Pesole, G. & Preparata, G. *Meth. Enzym.* **183**, 570–583 (1990).
18. Irwin, D. M., Kocher, T. D. & Wilson, A. C. *J. molec. Evol.* **32**, 128–144 (1991).
19. Amason, U., Gullberg, A. & Widegren, B. *J. molec. Evol.* **33**, 556–568 (1991).
20. Graur, D. & Higgins, D. G. *Molec. Biol. Evol.* **11**, 357–364 (1994).
21. Krettek, A., Gullberg, A. & Amason, U. *J. molec. Evol.* **41**, 952–957 (1995).
22. Saitou, N. & Nei, M. *Molec. Biol. Evol.* **4**, 406–425 (1987).
23. Amason, U., Xu, X. & Gullberg, A. *J. molec. Evol.* **42**, 145–152 (1996).
24. Amason, U. & Johnsson, E. *J. molec. Evol.* **34**, 493–505 (1992).
25. Amason, U., Gullberg, A., Johnsson, E. & Ledje, C. *J. molec. Evol.* **37**, 323–330 (1993).
26. Xu, X. & Amason, U. *Gene* **148**, 357–362 (1994).
27. Devereux, J., Haeblerli, P. & Smithies, O. *Nucleic Acids Res.* **12**, 387–395 (1984).
28. Felsenstein, J. *PHYLP (Phylogeny Inference Package)*, Version 3.5c (University of Washington, Seattle, 1993).

ACKNOWLEDGEMENTS. This work was supported by MURST (Italy), the Swedish Natural Sciences Research Council, and an EEC grant.

CORRESPONDENCE and requests for materials should be addressed to C.S. (e-mail: saccone@area.ba.cnr.it). Protein and nucleotide alignments analysed in the study are available on the World-Wide Web at <http://www.ba.cnr.it/guineapi.html>.

Power laws governing epidemics in isolated populations

C. J. Rhodes & R. M. Anderson

Centre for the Epidemiology of Infectious Disease, Department of Zoology, University of Oxford, South Parks Road, Oxford OX1 3PS, UK

TEMPORAL changes in the incidence of measles virus infection within large urban communities in the developed world have been the focus of much discussion in the context of the identification and analysis of nonlinear and chaotic patterns in biological time series^{1–11}. In contrast, the measles records for small isolated island populations are highly irregular, because of frequent fade-outs of infection^{12–14}, and traditional analysis¹⁵ does not yield useful insight. Here we use measurements of the distribution of epidemic sizes and duration to show that regularities in the dynamics of such systems do become apparent. Specifically, these biological systems are characterized by well-defined power laws in a manner reminiscent of other nonlinear, spatially extended dynamical systems in the physical sciences^{16–19}. We further show that the observed power-law exponents are well described by a simple lattice-based model which reflects the social interaction between individual hosts.

Isolated island populations have often provided a valuable arena for the study of evolutionary and ecological processes because the subject of investigation is confined to a well-defined region of space that is well insulated from the influence of external factors. In the context of population studies, it is often possible in such settings to better observe effects that are intrinsic to the community without having to quantify the effect of peripheral populations.

The Faroe Islands (population 25,000) have extensive detailed measles-case returns (Fig. 1A, top). It is considered to be an accurate epidemiological data set because the population is small and localized. Also, because of the comparative rarity of measles outbreaks, few measles cases escape notice²⁰. In the 58-year interval, there are 43 distinct epidemic events. An epidemic event is defined as the presence of a finite number of cases recorded in a sequence of consecutive months bounded by an absence of cases. An epidemic event has a duration, t , where $t = \tau_{\text{end}} - \tau_{\text{start}}$, τ_{start} is the month when cases in an event first appear, and τ_{end} is the next month when there are no more cases present. Similarly, an epidemic event has a size, s , where $s = \sum_{\tau=\tau_{\text{start}}}^{\tau_{\text{end}}} C(\tau)$, and $C(\tau)$ is the number of recorded cases of measles in the month τ .

From Fig. 1A, top, it is apparent that there is no obvious discernible pattern to the dynamics; there is wide variation in

FIG. 1 A, Faroe Islands; B, Bornholm; C, Reykjavik (scaled). These are all examples of Bartlett's type III measles epidemic dynamics¹². Top, monthly measles-case returns for each of the three communities. Between 1920 and 1970 the population of Reykjavik grew from 20,000 to 100,000, so we have scaled the incidence data per 25,000 of population. Middle, the distribution of epidemic event sizes. For the Faroes, $b \approx 0.28$; for Bornholm, $b \approx 0.28$; and for Reykjavik, $b \approx 0.21$. Bottom, distribution of epidemic event durations. For the Faroe Island data, $c \approx 0.8$; for Bornholm, $c \approx 0.85$; and for Reykjavik, $c \approx 0.62$. Clearly, scaling the measles incidence data will affect the epidemic sizes but leave the epidemic durations unscaled, hence this exponent is somewhat different from the Faroe and Bornholm cases. For each population, the exponents are unaffected by the specific choice of binning by decade.

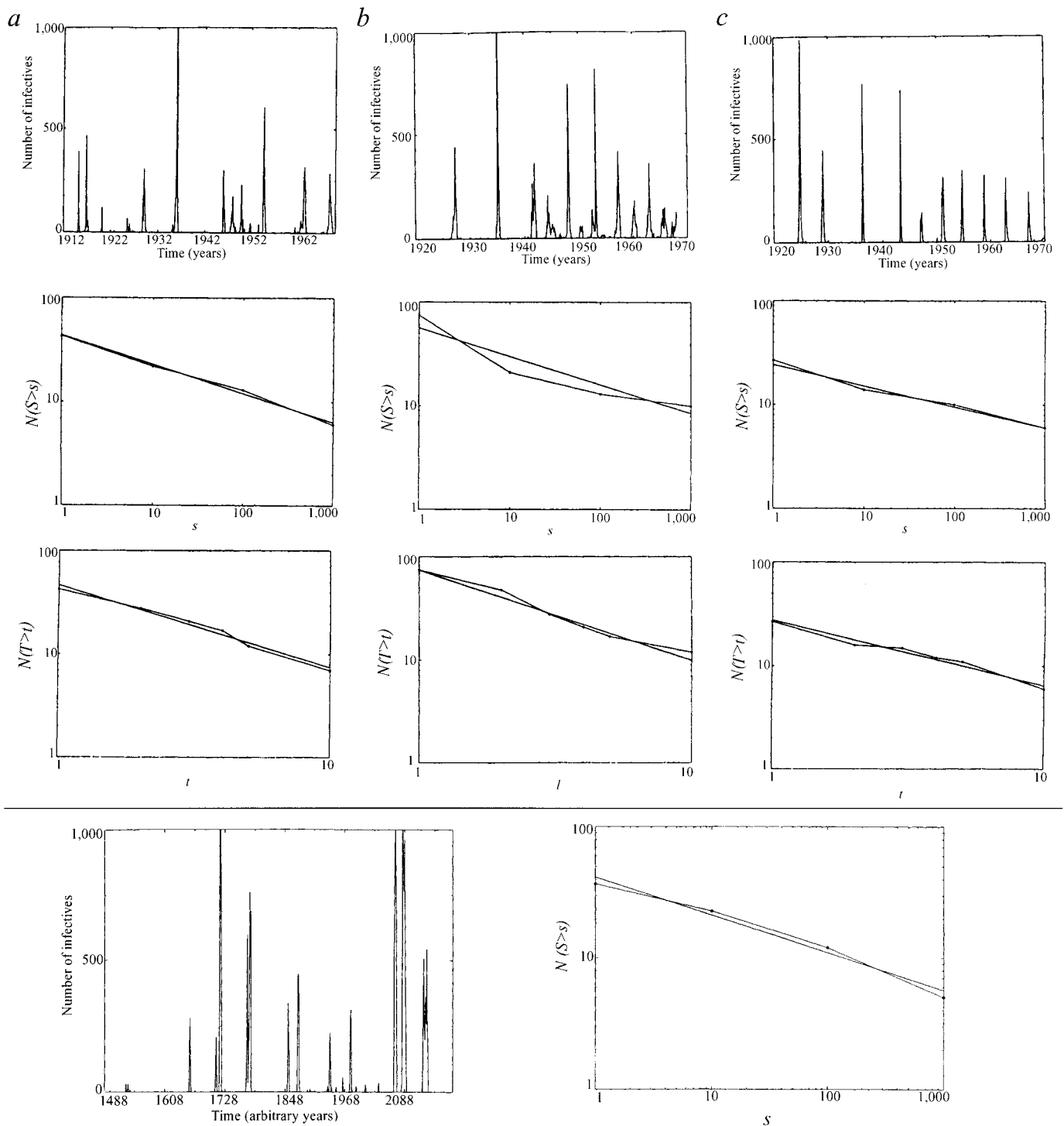
both the duration and size of the epidemic events, reminiscent of the sort of dynamics observed in the study of earthquakes. In that case, the Gutenberg–Richter²¹ law acts as an organizing principle connecting the frequency and magnitude of larger and smaller earthquakes. Performing the same analysis here, counting the number of epidemic events of size $> s$, there is a power-law dependence of the form $\log N(> s) = a - b \log(s)$. Thus, the number of epidemic events of size s scales as $N(s) \propto s^{-1-b}$, where $b \approx 0.28$. The same analysis for the epidemic event durations indicates that the number of epidemics of duration t scales as $N(t) \propto t^{-1-c}$, where, from Fig. 1A, bottom, $c \approx 0.8$. The presence of these scaling relations provides a useful way of estimating the likelihood, in a given interval of time, of epidemics of size s and duration t . Small short epidemics occur more often than large long epidemics, although their occurrence is connected and governed by the scaling exponents b and c . The exponents also apply to subportions of the data so, for example, we can use the epidemic distribution measured from the first half of the data set to predict the likely distribution of epidemic events in the second half.

We have also applied this analysis to two other subendemic populations with extensive and accurate measles records, Bornholm and Reykjavik. Figure 1B, C indicates that the exponents, b and c , for all three populations are remarkably similar.

Mechanisms leading to the emergence of power laws of this sort are still not fully understood, although recently many spatially extended interacting model systems have been shown to exhibit similar behaviour. We use one such lattice-based model, previously used in discussion of 'forest-fire' dynamics^{18,22,23}, to provide a simple model for this phenomenon.

The dynamics of the model in Fig. 2 closely resemble those seen in the Faroe data. Exponents for the model data are similar to the actual epidemiological data; $b \approx 0.29$ and $c \approx 1.5$, with the simulation underestimating the number of long epidemics of more than 10 months duration. The connectedness of the spatial distribution of the population on the lattice seems to reflect the social networks that exist in real communities²⁴.

FIG. 2 Top left, time series of infectives for the model simulation. Each site in the $L \times L$ lattice is in one of three states: empty, occupied by a susceptible, or occupied by an infective. The lattice is updated synchronously at each time-step using the following rules: (1) susceptibles who are on nearest-neighbour sites to an infective become infective themselves; (2) infectives become inactive and the site they occupy becomes empty; (3) susceptibles are introduced onto empty lattice sites with a probability μ ; (4) a new infective can arise when a susceptible is spontaneously infected with a probability ν . This effectively corresponds to an immigration term, whereby our lattice population is subject to infrequent infection from external sources. Periodic boundary conditions are used. These rules define a simple spatial S-I model. It is the simplest possible model of epidemic spread on a lattice and is, at best, a caricature of the known epidemiological processes. The lattice is 250×250 , with $\mu = 0.000026$ and $\nu = \mu/300$.



Mean population density on the lattice is 25,000. At equilibrium, with an average life expectancy of 70 years, we expect ~ 1 new susceptible to appear each day. The 43 epidemic events in 58 years sets a lower limit of $v/\mu \approx 1/400$, although we use $v/\mu \approx 1/300$ for the simulations. Birth of new susceptibles and immigration of infectives onto the lattice are modelled as Poisson processes, with a mean of 1 new susceptible appearing on the lattice at each time-step and a mean of one infective immigrating every 300 time-steps. Thus 1 time-step can be equated with one day. The x-axis is labelled for 696 arbitrary months (58 years). Top right, distribution of epidemic event sizes for the simulation data, $b \approx 0.29$. Bottom, distribution of epidemic event durations for the simulation data, $c \approx 1.5$. The exponent is severely affected by the poor prediction of long epidemics > 10 months, but comparison with Fig. 1A, bottom, shows the simulation makes a good prediction of the distribution of epidemics of up to 5 months duration for the Faroes.

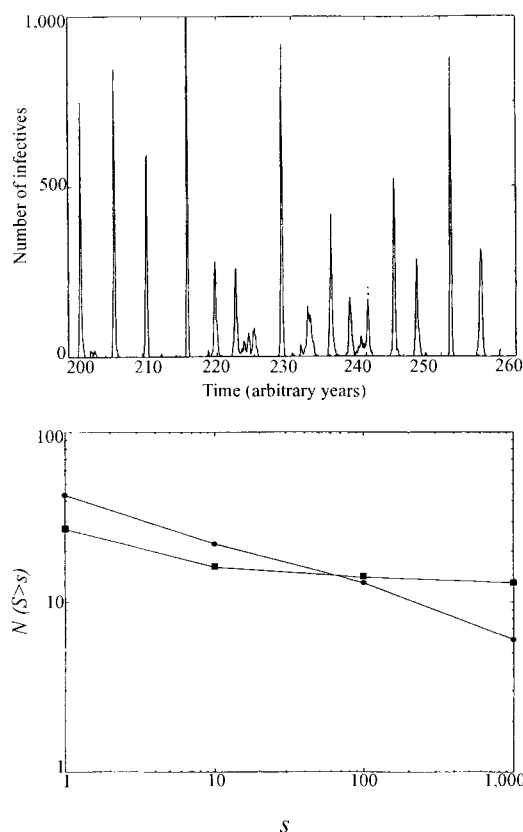


FIG. 3 Top left, time series of monthly number of infectives; bottom left, distribution of epidemic sizes (squares); and top right, epidemic duration (squares) for a stochastic Monte Carlo implementation of an SEIR model of measles in the Faroe Island population. The scaling for the Faroe Island (circles) data from Fig. 1 is shown for comparison. A constant population of 25,000 is used with a life expectancy of 50 years. The incubation and infectious intervals of measles are taken to be 7 days. R_0 is chosen to be 11. All births are into the susceptible class, with natural mortality occurring from each of the four classes. There is a Poisson influx of infectives into the community, on average every 300 days, as in the lattice simulation. The transients of the system are run off before data are recorded. No seasonality in the contact rate is used in these graphs, but it is found that inclusion of a sinusoidal seasonal driving term in the contact rate does not improve agreement between the model and the observed scaling.

For comparison, we illustrate the dynamics generated by a conventional stochastic SEIR model^{3,7,8} of the Faroe Islands. Conventional measles parameters are used and the population is modelled as a single homogeneously mixing community, with the same low-level immigration rate of external infectives. As before, we calculate the distribution of epidemic sizes and durations (Fig. 3). The resulting times series of infectives looks similar to the observed island dynamics, but this calculation overestimates the number of larger epidemics and is not able to fit the observed data as well as the lattice model. The assumption of homogeneous mixing of susceptibles with infectives that underlie the SEIR equations appear to break down for small populations with infrequent epidemics. The replenishment of susceptibles onto vacant lattice sites and the nearest-neighbour contact spread of infection are central to reproducing the observed epidemic

dynamics.

Our results suggest the existence of scaling laws for the size and duration of epidemics in the isolated island measles data sets. This places the dynamics of these epidemics in the same class as other spatially extended nonlinear dynamical systems, where scaling is also observed. In practice, this facilitates a form of prediction in which we can calculate the frequency of occurrence of epidemics of given size and duration. A simple spatial model generates almost the same exponents as are seen in our data analysis. The power-law phenomenon is also likely to be of relevance in the study of infrequent outbreaks of measles in highly vaccinated communities and in isolated rural populations in developing countries. The methods discussed here are completely general and could be applied to any other times series of communicable disease outbreaks in a small population. □

Received 9 February; accepted 11 April 1996.

- May, R. M. *Nature* **261**, 459–467 (1967).
- May, R. M. *Proc. R. Soc. Lond. A* **413**, 27–44 (1987).
- Grenfell, B. T. *J. R. statist. Soc. B* **54**, 383–398 (1992).
- Grenfell, B. T., Kleczkowski, A., Ellner, S. & Bolker, B. M. *Phil. Trans. R. Soc. Lond. A* **348**, 515–530 (1994).
- Bolker, B. M. & Grenfell, B. T. *Proc. R. Soc. Lond. B* **251**, 75–81 (1993).
- Grenfell, B. T., Bolker, B. M. & Kleczkowski, A. *Proc. R. Soc. Lond. B* **259**, 97–103 (1995).
- Olsen, L. F., Truty, G. L. & Schaffer, W. M. *Theor. Pop. Biol.* **33**, 344–370 (1988).
- Olsen, L. F. & Schaffer, W. M. *Science* **249**, 499–504 (1990).
- Sugihara, G. & May, R. M. *Nature* **344**, 734–741 (1990).
- Sugihara, G., Grenfell, B. & May, R. M. *Phil. Trans. R. Soc. Lond. B* **330**, 235–251 (1990).
- Tidd, C. W., Olsen, L. F. & Schaffer, W. M. *Proc. R. Soc. Lond. B* **254**, 257–273 (1993).
- Bartlett, M. S. *J. R. statist. Soc. A* **120**, 48–70 (1957).
- Bartlett, M. S. *J. R. statist. Soc. A* **123**, 37–44 (1960).
- Black, F. L. *J. theor. Biol.* **11**, 207–211 (1966).
- Anderson, R. M. & May, R. M. *Infectious Diseases of Humans, Dynamics and Control* (Oxford Univ. Press, 1991).
- Bak, P., Tang, C. & Wiesenfeld, K. S. *Phys. Rev. A* **38**, 364–374 (1988).

- Chen, K., Bak, P. & Obukhov, S. P. *Phys. Rev. A* **43**, 620–635 (1991).
- Clar, S., Drossel, B. & Schwabl, F. *Phys. Rev. E* **50**, 1009–1018 (1994).
- Sornette, A. & Sornette, D. *Europhys. Lett.* **9**, 197–202 (1989).
- Cliff, A. D., Haggett, P. & Smallman-Raynor, M. *Measles. An Historical Geography of a Major Human Viral Disease from Global Expansion to Local Retreat 1840–1990* (Blackwell, Oxford, 1993).
- Gutenberg, B. & Richter, C. F. *Ann. Geofis.* **9**, 1–10 (1956).
- Bak, P., Chen, K. & Tang, C. *Phys. Lett. A* **147**, 297–300 (1990).
- Drossel, B. & Schwabl, F. *Phys. Rev. Lett.* **69**, 1629–1632 (1992).
- Ferguson, N., May, R. M. & Anderson, R. M. in *Spatial Ecology: The Role of Space in Population Dynamics and Interspecific Interactions* (eds Tilman, D. & Kareiva, P.) (Princeton Monographs in Population Biology, in the press).

ACKNOWLEDGEMENTS. We thank L. F. Olsen of Odense University for use of the Faroe Island and Bornholm measles data, A. D. Cliff of Cambridge University for use of the Iceland and Reykjavik measles data, and N. Ferguson of Oxford University for many useful discussions. Research support for C.J.R. and R.M.A. is provided by the Wellcome Trust.

CORRESPONDENCE should be addressed to C.J.R. (e-mail: chris.rhodes@zoology.oxford.ac.uk).

Role of transcription factors Brn-3.1 and Brn-3.2 in auditory and visual system development

Linda Erkman*, Robert J. McEvilly*, Lin Luo†, Aimee K. Ryan*, Farideh Hooshmand*, Shawn M. O'Connell*, Elizabeth M. Keithley†, David H. Rapaport§, Allen F. Ryan† & Michael G. Rosenfeld*

* Howard Hughes Medical Institute, Department and School of Medicine, § Department of Surgery, Division of Anatomy, † VA Medical Center and Departments of Surgery/Otolaryngology and Neurosciences, University of California, San Diego, 9500 Gilman Drive, La Jolla, California 92093, USA

THE neurally expressed genes *Brn-3.1* and *Brn-3.2* (refs 1–6) are mammalian orthologues of the *Caenorhabditis elegans unc-86* gene⁷ that constitute, with *Brn-3.0* (refs 1–3,8,9), the class IV POU-domain transcription factors¹⁰. *Brn-3.1* and *Brn-3.2* provide a means of exploring the potentially distinct biological functions of expanded gene families in neural development. The highly related members of the Brn-3 family have similar DNA-binding preferences^{1,2} and overlapping expression patterns in the sensory nervous system, midbrain and hindbrain^{1–6,8,9}, suggesting functional redundancy. Here we report that *Brn-3.1* and *Brn-3.2* critically modulate the terminal differentiation of distinct sensorineural cells in which they exhibit selective spatial and temporal expression patterns. Deletion of the *Brn-3.2* gene causes the loss of most retinal ganglion cells, defining distinct ganglion cell populations. Mutation of *Brn-3.1* results in complete deafness, owing to a failure of hair cells to appear in the inner ear, with subsequent loss of cochlear and vestibular ganglia.

Detailed investigation of class IV POU domain factor expression indicates that, in addition to extensive overlap in the nervous system, each gene displays unique spatio-temporal patterns, suggesting that they are important for the development of specific populations of neurons. For example, *Brn-3.0* and *Brn-3.2* exhibit distinct temporal patterns of expression in retinal ganglion cells. *Brn-3.2* is initially detected at embryonic day 13.5 (E13.5), peaks at E15.5 when most of the retinal ganglion cells have been generated^{11,12}, declines later in development, and is almost undetectable by postnatal day 8 (P8) (Fig. 1a). In contrast, expression of the *Brn-3.0* gene has a later onset (at E15.5) and exhibits a slower decrease, with moderate levels of protein still detectable at P8 (Fig. 1a).

To investigate the role of *Brn-3.2* in neural development, we have deleted the entire protein-coding region of the *Brn-3.2* gene by targeted mutagenesis (Fig. 1b–e). Mice heterozygous and homozygous for *Brn-3.2* deletion do not display obvious behavioural

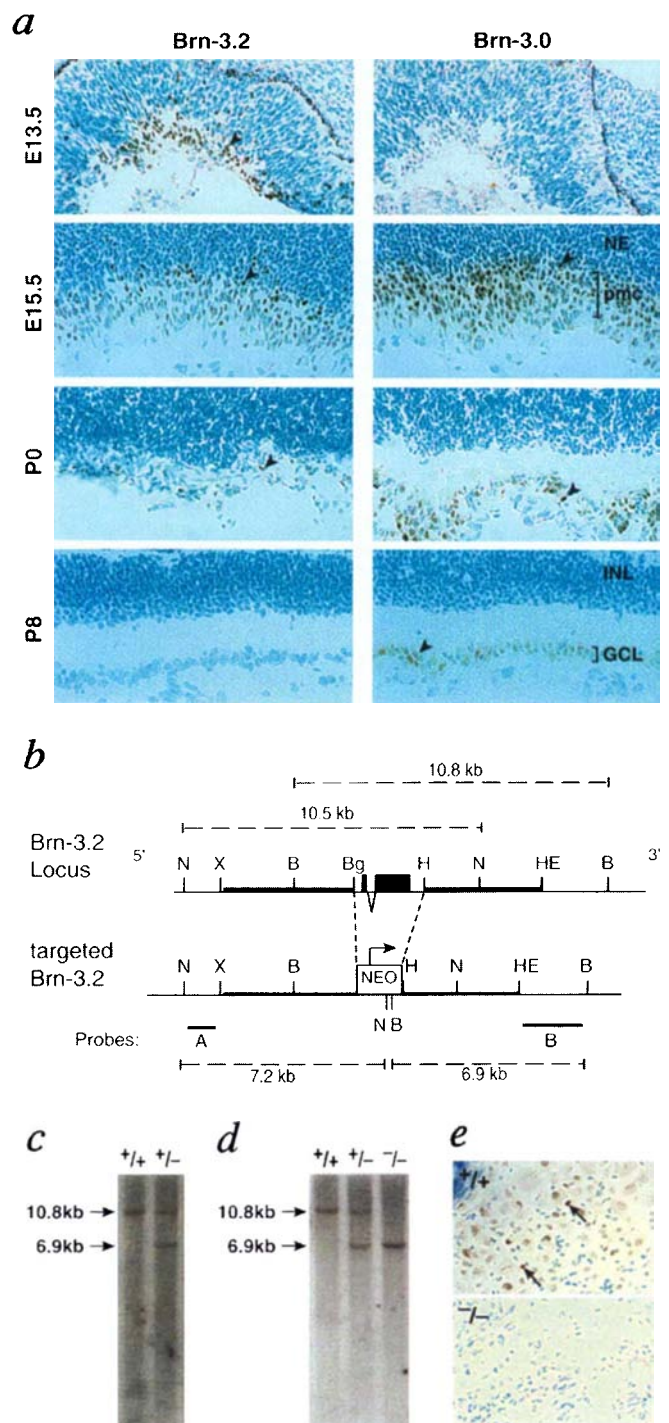


FIG. 1 Generation of *Brn-3.2* gene-deleted mice. **a**, Ontogeny of *Brn-3.2* and *Brn-3.0* expression in E13.5, E15.5, P0 and P8 retinas as assessed by immunohistochemistry. Arrowheads show nuclear staining in ganglion cells. GCL, ganglion cell layer; INL, inner nuclear layer; NE, neuro-epithelium; pmc, zone of postmitotic cells. **b**, Diagram of the *Brn-3.2* gene, the disrupted *Brn-3.2* allele, and diagnostic probes (A and B). Filled boxes represent the two coding exons, and thick bars homologous sequences used in the targeting vector. Broken lines indicate the predicted size of restriction fragments recognized by probes in the wild-type and mutant allele, respectively. In the targeting vector the positive selection marker (Neo) replaced 2.7 kb of the genomic locus between the *Bgl*III and *Hind*III sites, and two negative selection markers, HSV-1 TK gene under the control of *pgk*-1 promoter, and HSV-2 TK gene driven by *pmc*-1 promoter, were inserted at the 5' and 3' ends of homologous sequences. B, *Bam*HI; Bg, *Bgl*III; E, *Eco*RI; H, *Hind*III; N, *Nco*I; X, *Xho*I. **c**, Southern blot analysis of electroporated embryonic stem (ES) cell genomic DNA with probe B

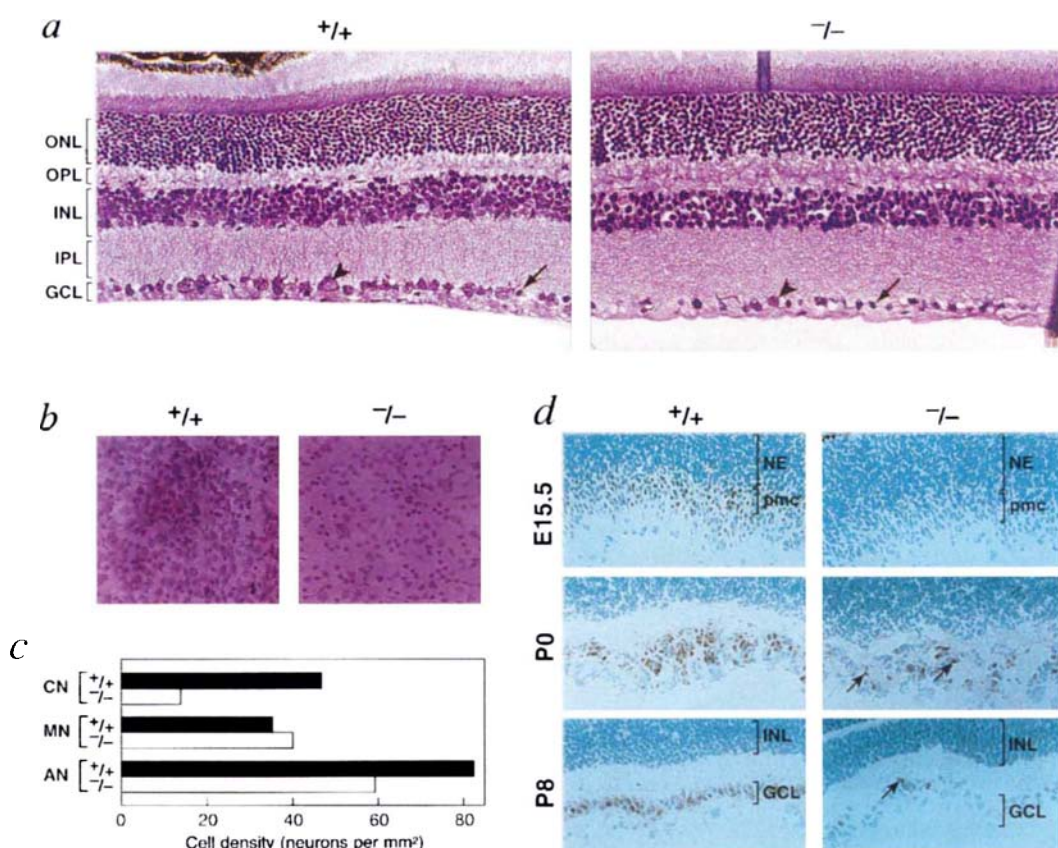
following *Bam*HI digestion yields the predicted size for wild-type (10.8 kb) and mutant (6.9 kb) alleles. **d**, Restriction fragment length polymorphism (RFLP) analysis of an intercross litter from mice heterozygous for the *Brn-3.2* null allele. **e**, Analysis of *Brn-3.2* protein expression in *Brn-3.2*^{+/+} and *Brn-3.2*^{-/-} P0 dorsal root ganglia using anti-*Brn-3.2* antiserum. Arrows show positive nuclei.

METHODS. Anti-*Brn-3.0* and anti-*Brn-3.2* immunohistochemistry. For antibody preparation, rabbit antisera were raised against bacterially expressed fusion peptides encompassing amino acids 189–264 of *Brn-3.0* protein, and amino acids 185–254 of *Brn-3.2*, as described²¹. Immunohistochemistry was performed on fresh-frozen cryostat sections from *Brn-3.2*^{+/+} and *Brn-3.2*^{-/-} littermates, fixed for 10 min in 4% paraformaldehyde at 4 °C, before incubation with primary antibodies. The reaction was revealed with diaminobenzidine following incubation with an anti-rabbit antibody coupled to peroxidase (Amersham). The sections were counterstained with methyl green. Targeted mutagenesis was performed as previously described^{21,22}.

defects, develop to adulthood, and are fertile. The main phenotype of *Brn-3.2*^{-/-} mice is observed in the eye. Although the overall structure is preserved, mature (P17–P120) retinal whole mounts from *Brn-3.2*^{-/-} mice show a loss of approximately 60–70% of retinal ganglion cells (Fig. 2a–c). Counts of profiles in the ganglion cell layer of neonatal (P0) retina sections and whole mounts are similar for the wild type and mutant (data not shown), suggesting that the loss of ganglion cells occurs postnatally.

Deletion of *Brn-3.2* also affects *Brn-3.0* expression in the retina, with *Brn-3.0* protein failing to appear at its normal time of onset in the E15.5 retina of mutant mice (Fig. 2d). Although we cannot exclude the possibility that *Brn-3.2* directly regulates *Brn-3.0* expression in a subpopulation of retinal ganglion cells, this appears unlikely because *Brn-3.2* mutation does not result in failure to express *Brn-3.0* in other regions of the nervous system (data not shown). Based on the initial expression of *Brn-3.2* in postmitotic cells in the central nervous system (CNS)¹³, and the absence of a difference in the number of ganglion cells in P0 retina, it is equally unlikely that the absence of *Brn-3.0* expression at E15.5 reflects an early loss of ganglion cell progenitors by failure to proliferate or premature death. Therefore, we hypothesize that the ultimate loss of ganglion cells in *Brn-3.2*^{-/-} mice results from a failure of proper differentiation and gene activation, impairing the ability of these cells to respond to the trophic factors and/or environmental cues required for the establishment of synaptic connections with CNS targets. The few ganglion cells that subsequently express *Brn-3.0* in P0 and P8 mutants (Fig. 2d) probably represent a late *Brn-3.0*-expressing population that develops independently of *Brn-3.2*. This suggests that, despite extensive overlap of the two genes in ganglion cells, some '*Brn-3.0* only' cells exist in the mouse retina.

FIG. 2 Loss of retinal ganglion cells in *Brn-3.2* gene-deleted mice. a, b, Sections and whole mounts of comparable regions of retinas of *Brn-3.2*^{-/-} and *Brn-3.2*^{+/+} littermates, exhibiting a paucity of retinal ganglion cells in *Brn-3.2*^{-/-} mouse retina. Neuronal profiles in the ganglion cell layer were subdivided, based on morphological and staining differences, into 'classical neurons' (ganglion cells and large, displaced amacrine cells; arrowheads), and 'microneurons' (most displaced amacrine cells^{23–25}; small arrows). GCL, ganglion cell layer; INL, inner nuclear layer; IPL, inner plexiform layer; ONL, outer nuclear layer; OPL, outer plexiform layer. c, Histogram of cell density in the ganglion cell layer measured from whole mounts of wild-type and *Brn-3.2*^{-/-} mice. Overall there are 35–40% fewer neuronal profiles (all neurons; AN) in the ganglion cell layer of *Brn-3.2*^{-/-} mice compared to *Brn-3.2*^{+/+} mice; the 'classical neuron' (CN) density is 60–70% less than in the wild type. In contrast, there are more microneurons (MN) in the ganglion cell layer of the mutant. d, Expression of *Brn-3.0* protein in wild-type and mutant retinas at E15.5, P0 and P8. *Brn-3.0* expression is not detectable in *Brn-3.2*^{-/-} mice at E15.5. Arrows indicate *Brn-3.0*-positive nuclei in P0 and P8 mutant retina. NE, neuroepithelium; pmc, zone of postmitotic cells; INL, inner nuclear layer; GCL, ganglion cell layer. METHODS. Plastic sections and whole mounts of the retinas were prepared



Brn-3.2 is also expressed at high levels in the superior colliculus², the major target of retinal ganglion cells. Histological analysis suggests disorganized morphology of the superior colliculus, even though *Brn-3.0* expression, for example, is not disturbed (data not shown). Specifically, there are subtle differences in the cell density and the laminar organization that are particularly apparent in the stratum zonale (data not shown). The limited phenotypic abnormalities in the colliculus of *Brn-3.2*^{-/-} mice could reflect either the compensatory actions of *Brn-3.0* and/or the persistence of a residual ganglion-cell input, as well as the involvement of these structures with neural systems other than the retina. The central nervous system of *Brn-3.2*^{-/-} mice appears normal in other respects; examination from the spinal cord to the rostral-most limit of *Brn-3.2* expression in the midbrain using serial paraffin sections and *in situ* hybridization with probes for neuropeptides, neurotrophin receptors and transcription factors reveals no gross abnormalities (data not shown). Similarly, peripheral sensory ganglia do not display an overt phenotype, probably because of extensive overlap and functional redundancy with *Brn-3.0*. Therefore, the striking developmental consequence of *Brn-3.2* gene deletion is the loss of a large population of retinal ganglion cells. This observation suggests the existence of two distinct regulatory pathways that lead to the formation of *Brn-3.2*-dependent and -independent ganglion cells.

To explore whether *Brn-3.1*, the closely related orthologue, might also perform a distinct, restricted role in neural development, we examined the ontogeny and spatial aspects of its expression. In trigeminal and dorsal root ganglia, developing brain, retina and some scattered cells in the dorsal horn of the spinal cord, *Brn-3.1* expression overlaps with that of *Brn-3.0* and *Brn-3.2* (refs 1,3,6, and our unpublished data). However, *Brn-3.1* is uniquely and

as described^{26,27}, and counts were performed on wild-type and mutant retinas from littermates on pure 129/Sv genetic background. Corresponding areas of wild-type and mutant littermates were viewed with a X100 objective lens, and the number of profiles in ten contiguous 100- μ m² sample regions were counted. *Brn-3.0* immunohistochemistry was performed as described in Fig. 1a.

strongly expressed in cochlear and vestibular hair cells of the inner ear (Fig. 3a), and is expressed in the vestibular labyrinth by E14 (data not shown), in the cochlea by E18, and continues into adulthood (Fig. 3a). Expression appears limited to hair cells, as supporting cells beneath and adjacent to them are not labelled.

A null mutation in the *Brn-3.1* locus was created by replacing the normal gene with a disrupted version lacking the entire protein-coding region (Fig. 3b–e). Heterozygous mice appear normal in all regards and are fertile. *Brn-3.1*^{-/-} mice show deficits in balance and coordination that become severe at about P14, and exhibit marked hyperactivity by 5–6 weeks. Deletion of the *Brn-3.1* gene results in complete deafness. At P42, both *Brn-3.1*^{+/+} and *Brn-3.1*^{+/-} mice have normal click-evoked auditory brainstem responses, whereas *Brn-3.1*^{-/-} mice show no evoked potentials to stimuli up to 80 dB above normal threshold for the mouse (Fig. 3f). Mouse–human synteny (ref. 14 and unpublished data) indicates that the *Brn-3.1* genomic locus is on human chromosome 5q21–35 and encompasses the locus 5q31, where an inherited

form of late-onset degenerative deafness (after age 30) has been mapped¹⁵. Thus a mutation in the *Brn-3.1* genomic locus represents a strong candidate for this form of genetic deafness.

To study the cellular basis of this physiological defect, we examined the morphology of cochlear and vestibular hair cells at P0 and P14. The cochlea and spiral ganglia of *Brn-3.1*^{-/-} mice appear grossly normal at P0 (Fig. 4a). However, in the organ of Corti, cells at the normal location of the inner and outer hair cells fail to show early signs of differentiation, including growth of stereocilia and clear segregation of cell nuclei to a plane above the level of supporting cell nuclei. By P14, the defect in the organ of Corti has progressed, and neither hair cells nor many of the normal supporting cell populations can be identified. Most spiral ganglion cells have degenerated (Fig. 4b), and the vestibular labyrinth is equally affected. At P0, the gross appearance of the vestibular labyrinth of *Brn-3.1*^{-/-} mice is normal, although there is a specific lack of recognizable hair cells in the sensory epithelia of the maculae and cristae (Fig. 4c). No stereociliary bundles are

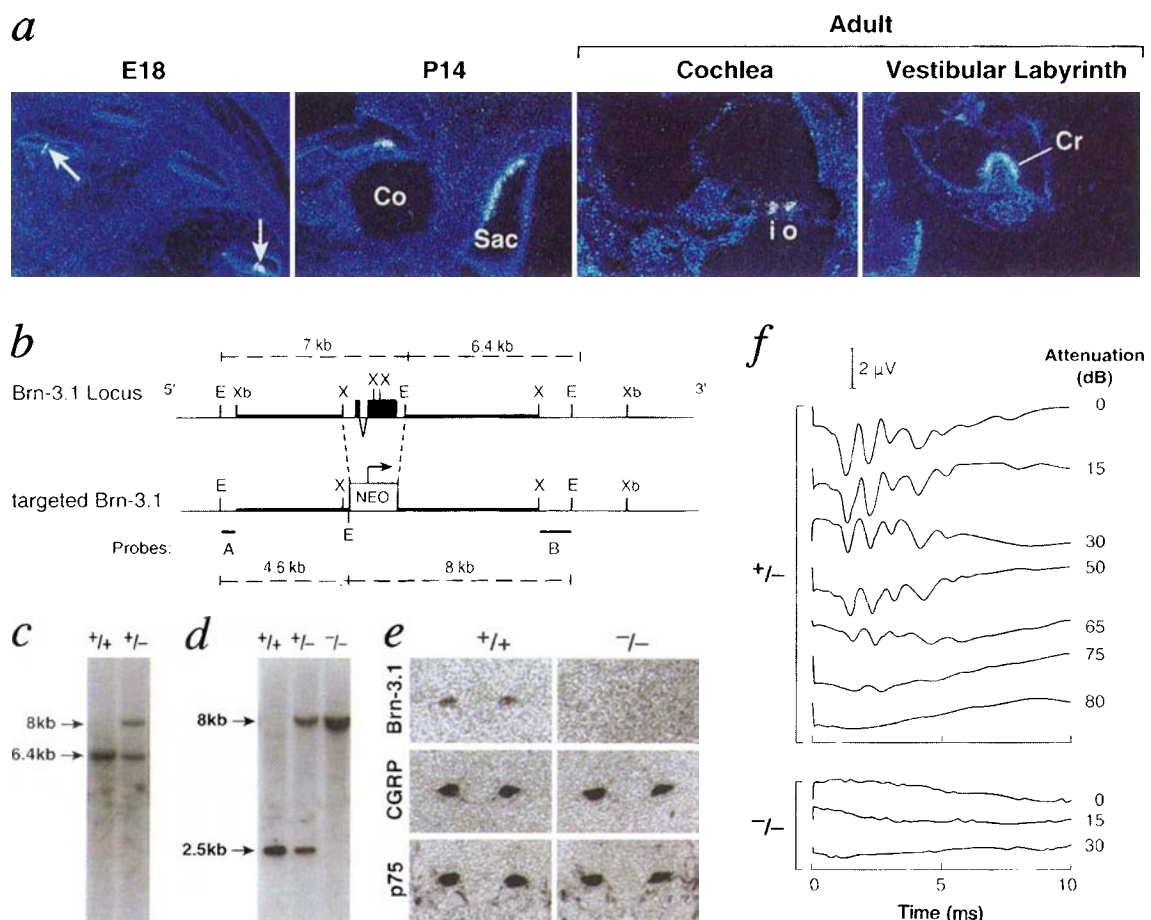


FIG. 3 Generation of *Brn-3.1* mutant mice. **a**, Expression of *Brn-3.1* mRNA in developing rat cochlea. Expression begins in the basal half of the cochlea, which matures earlier than the apex. High levels of expression are maintained in adult hair cells. Co, cochlea; Cr, Crista ampularis; i, inner hair cell; o, outer hair cell; Sac, saccule. **b**, *Brn-3.1* genomic locus, disrupted allele and diagnostic probes (A and B). For targeting strategy see Fig. 1b. Boxes and thick bars represent exons and homologous sequences used in the targeting vector, respectively. The Neomycin resistance gene replaces 2.4 kb of the genomic locus between the *Xho*I and *Eco*RI sites, disrupting the entire protein coding region. E, *Eco*RI; X, *Xho*I; Xb, *Xba*I. **c**, RFLP analysis of ES cell lines produces the expected wild-type (6.4 kb) and mutant (8 kb) *Eco*RI fragments, whereas in **d**, analysis of mouse genomic DNA yields the expected 8-kb mutant band, and a wild-type band of 2.5 kb. The difference in the wild-type *Eco*RI band between the ES cell lines (of 129/Sv origin) and the mice (C57/BL6) is due to a polymorphism at the *Brn-3.1* locus. **e**, *In situ* hybridization analysis of *Brn-3.1*, p75 low-affinity NGFR, and α -CGRP mRNA expression in the trigeminal ganglia of *Brn-3.1*^{+/+} and *Brn-3.1*^{-/-} littermates in coronal sections of the whole heads. **f**, Deafness in *Brn-3.1* mutant mice. *Brn-3.1*^{-/-} mice showed a complete failure of click-evoked auditory brainstem response.

METHODS. *In situ* hybridization on the inner ear of embryonic and adult rats was performed as described²⁸, using a *Brn-3.1* cRNA probe corresponding to amino acids 242–311. Targeted mutagenesis was performed as described in Fig. 1b. *In situ* hybridization with *Brn-3.1*, p75 low-affinity NGFR and α -CGRP probes was done as described²⁹. cRNA probes were derived from the third exon of mouse p75 cDNA and from the fifth and sixth exons of mouse α -CGRP. Hearing tests were performed on anaesthetized animals (0.05 mg per g ketamine; 0.005 mg per g acepromazine; 0.001 mg per g rompun, intramuscular) using a Tucker-Davis computer-controlled system. A biphasic click (100 μ s 10 per s) was presented to a Beyer DT 48 earphone rigidly coupled to the animal's external meatus, and evoked potentials ($n = 512$) were recorded from subcutaneous needle electrodes placed postauricularly and on the vertex (gain = 10,000). Stimuli were presented in decreasing amplitude steps until threshold was reached.

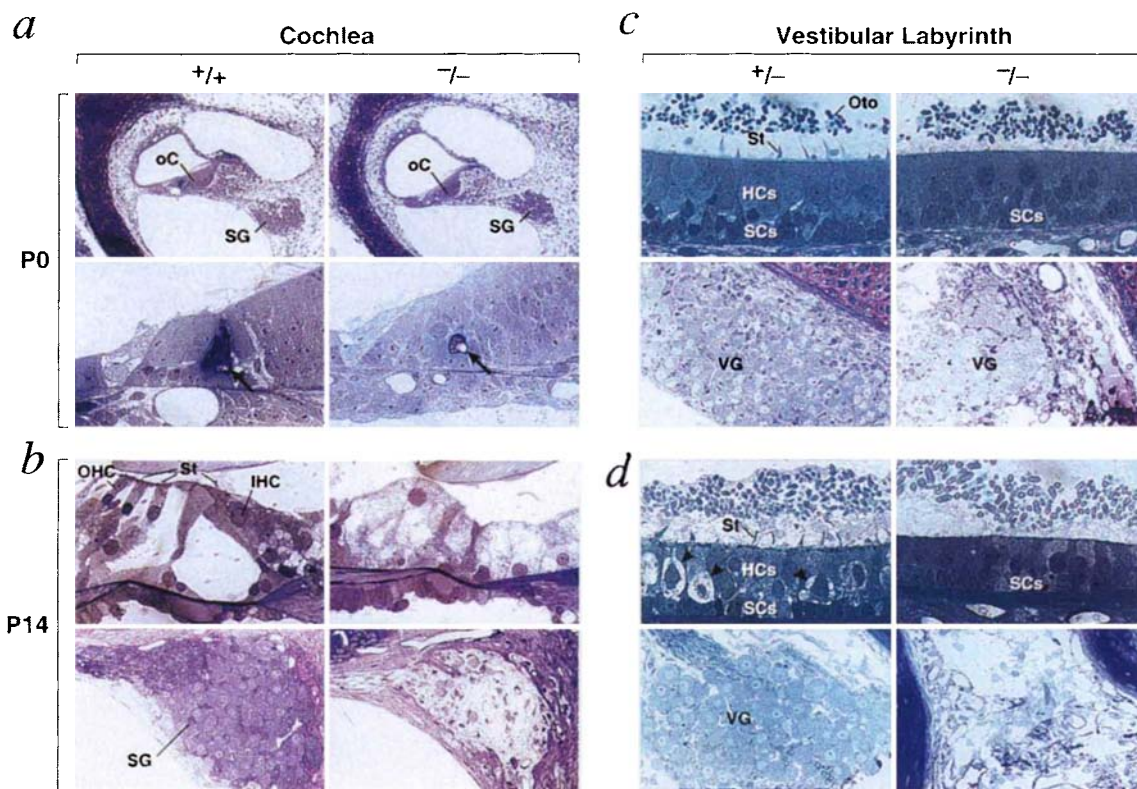


FIG. 4 Loss of hair cells and primary sensory neurons in *Brn-3.1* mutant mice. Morphology of the cochlea (a, b), and vestibular labyrinth (c, d), in neonatal (P0) (a, c) and P14 (b, d) wild-type and mutant mice. Note the absence of stereocilia, abnormal morphology of cells in the hair cell region, and degeneration of sensory ganglion neurons. Normal development of the organ of Corti is further disrupted in *Brn-3.1* mutant mice, beginning with incomplete cavitation of the tunnel of Corti at P0 (arrows), and including

the absence or severe distortion of supporting cells at P14. Calycial endings surrounding type I hair cells (arrowheads) are absent from the mutant vestibular system at P14. IHC, inner hair cell; HCs, hair cells; oC, organ of Corti; OHC, outer hair cell; Oto, otoliths; SCs, supporting cells; SG, spiral ganglion; St, stereocilia bundles; VG, vestibular ganglion. METHODS. Morphological assessment of the inner ear was performed as described³⁰.

present, even at the electron-microscopic level (data not shown), and the row of nuclei above the supporting cell nuclei is sparse and poorly organized. The vestibular ganglion shows evidence of neuronal degeneration, which is complete by P14, including absence of the calycial endings surrounding type I hair cells (Fig. 4d). The ontogeny of the organ of Corti and the survival of inner-ear sensory ganglia is thought to depend on the interaction of hair cells and supporting cells, involving lateral inhibition and trophic influences^{16–19}. Because *Brn-3.1* expression is apparently limited to hair cells, changes in the supporting cells of the organ of Corti and in spiral and vestibular ganglia of *Brn-3.1*^{-/-} mice

appear to be secondary to failure of hair-cell differentiation, providing additional evidence that hair cells play a critical role in the organization of the organ of Corti.

Taken together, our data indicate that both *Brn-3.1* and *Brn-3.2* critically affect terminal differentiation events in specific sensorineural cells, consistent with the role of *unc-86* in determining mechanosensory lineages in *C. elegans*²⁰. This suggests that expansion of these *unc-86* orthologues has been exploited in mammals for independent regulation of critical developmental events in two specialized sensory organs required for recognition of visual and auditory stimuli. □

Received 11 March; accepted 20 May 1996.

- Gerrero, M. R. et al. *Proc. natn. Acad. Sci. U.S.A.* **90**, 10841–10845 (1993).
- Turner, E. E., Jenne, K. J. & Rosenfeld, M. G. *Neuron* **12**, 205–218 (1994).
- Xiang, M. et al. *J. Neurosci.* **15**, 4762–4785 (1995).
- Xiang, M. et al. *Neuron* **11**, 689–701 (1993).
- Lillycrop, K. A. et al. *Nucleic Acids Res.* **20**, 5093–5096 (1992).
- Ninkina, N. N., Stevens, G. E., Wood, J. N. & Richardson, W. D. *Nucleic Acids Res.* **21**, 3175–3182 (1993).
- Finney, M., Ruvkun, G. & Horvitz, H. R. *Cell* **55**, 757–769 (1988).
- He, X. et al. *Nature* **340**, 35–42 (1989).
- Collum, R. G. et al. *Nucleic Acids Res.* **20**, 4919–4925 (1992).
- Wegner, M., Drolet, D. W. & Rosenfeld, M. G. *Curr. Opin. Cell Biol.* **5**, 488–498 (1993).
- Sidman, R. L. in *The Structure of the Eye* (ed. Smelser, G. K.) 487–506 (Academic, New York, 1961).
- Young, R. W. *Anat. Rec.* **212**, 199–205 (1985).
- Fedtsova, N. & Turner, E. E. *Mech. Dev.* **53**, 291–304 (1995).
- Xia, Y.-R. et al. *Genomics* **18**, 126–130 (1993).
- Leon, P. E., Raventos, H., Lynch, E., Morrow, J. & King, M.-C. *Proc. natn. Acad. Sci. U.S.A.* **89**, 5181–5184 (1992).
- Kelley, M. W., Xu, X.-M., Wagner, M. A., Warchol, M. E. & Corwin, J. T. *Development* **119**, 1041–1055 (1993).
- Ernfors, P. & Lee, K.-F. *Nature* **368**, 147–150 (1994).
- Ernfors, P., Van De Water, T., Loring, J. & Jaenisch, R. *Neuron* **14**, 1153–1164 (1995).
- Fannas, I., Jones, K. R., Backus, C., Wang, X.-Y. & Reichardt, L. F. *Nature* **369**, 658–661 (1994).
- Finney, M. & Ruvkun, G. *Cell* **63**, 895–905 (1990).

- Schonemann, M. D. et al. *Genes Dev.* **9**, 3122–3135 (1995).
- Mansour, S. L., Thomas, K. R. & Capecchi, M. R. *Nature* **336**, 348–352 (1988).
- Dräger, U. C. & Olsen, J. F. *Invest. Ophthalmol. vis. Sci.* **20**, 285–293 (1981).
- Wäslie, H., Chun, M. H. & Müller, F. J. *comp. Neurol.* **265**, 391–408 (1987).
- Wong, R. O. & Hughes, A. J. *comp. Neurol.* **255**, 159–177 (1987).
- Rapaport, D. H. & Vietri, A. J. *comp. Neurol.* **312**, 341–352 (1991).
- Stone, J. *The Wholemount Handbook* (Maitland, Sydney, 1981).
- Luo, L., Koutnouyan, H., Baird, A. & Ryan, A. F. *Hearing Res.* **69**, 182–193 (1993).
- Simmons, D. M. et al. *Genes Dev.* **4**, 695–711 (1990).
- Crenshaw, E. B., Ryan, A., Dillon, S. R., Kalia, K. & Rosenfeld, M. G. *J. Neurosci.* **11**, 1524–1530 (1991).

ACKNOWLEDGEMENTS. We thank M. Ayers and A. Uribe for assistance in the development of founder mice and maintenance of the animal facility; P. Sawchenko for examining the hindbrains of *Brn-3.2* gene-deleted mice; J. Bermingham for advice on embryonic stem cells; M. Schonemann for help with electroporation of the embryonic stem cells; T. Bennett for assistance with histology; E. Li for J1 embryonic stem cells; J. Huber for the p75 cRNA template; B. Andersen, A. Gleiberman and M. Treier for reading the manuscript; P. Myer for assistance with preparation of illustrations; and B. Stawski for help with manuscript preparation. A.K.R. is a recipient of a NRSA from the NIH. L.E. was funded by a postdoctoral fellowship from the Swiss National Science Foundation and the Ciba Foundation. L. Luo is supported by the House Ear Institute of Los Angeles. M.G.R. is an investigator with the Howard Hughes Medical Institute. This research was supported by grants from the NIH to M.G.R., and NIH/NIDCD and VA Research Service to A.F.R.

CORRESPONDENCE and requests for materials should be addressed to M.G.R. (e-mail: Mrosenfeld@ucsd.edu).

Emergence of simple-cell receptive field properties by learning a sparse code for natural images

Bruno A. Olshausen* & David J. Field

Department of Psychology, Uris Hall, Cornell University, Ithaca, New York 14853, USA

THE receptive fields of simple cells in mammalian primary visual cortex can be characterized as being spatially localized, oriented¹⁻⁴ and bandpass (selective to structure at different spatial scales), comparable to the basis functions of wavelet transforms^{5,6}. One approach to understanding such response properties of visual neurons has been to consider their relationship to the statistical structure of natural images in terms of efficient coding⁷⁻¹². Along these lines, a number of studies have attempted to train unsupervised learning algorithms on natural images in the hope of developing receptive fields with similar properties¹³⁻¹⁸, but none has succeeded in producing a full set that spans the image space and contains all three of the above properties. Here we investigate the proposal^{8,12} that a coding strategy that maximizes sparseness is sufficient to account for these properties. We show that a learning algorithm that attempts to find sparse linear codes for natural scenes will develop a complete family of localized, oriented, bandpass receptive fields, similar to those found in the primary visual cortex. The resulting sparse image code provides a more efficient representation for later stages of processing because it possesses a higher degree of statistical independence among its outputs.

We start with the basic assumption that an image, $I(x, y)$, can be represented in terms of a linear superposition of (not necessarily orthogonal) basis functions, $\phi_i(x, y)$:

$$I(x, y) = \sum_i a_i \phi_i(x, y) \quad (1)$$

The image code is determined by the choice of basis functions, ϕ_i . The coefficients, a_i , are dynamic variables that change from one image to the next. The goal of efficient coding is to find a set of ϕ_i that forms a complete code (that is, spans the image space) and results in the coefficient values being as statistically independent as possible over an ensemble of natural images. The reasons for desiring statistical independence have been elaborated elsewhere^{9,12,19}, but can be summarized briefly as providing a strategy for extracting the intrinsic structure in sensory signals.

One line of approach to this problem is based on principal-components analysis^{14,15,20}, in which the goal is to find a set of mutually orthogonal basis functions that capture the directions of maximum variance in the data and for which the coefficients are pairwise decorrelated, $\langle a_i a_j \rangle = \langle a_i \rangle \langle a_j \rangle$. The receptive fields that result from this process are not localized, however, and the vast majority do not at all resemble any known cortical receptive fields (Fig. 1). Principal components analysis is appropriate for capturing the structure of data that are well described by a gaussian cloud, or in which the linear pairwise correlations are the most important form of statistical dependence in the data. But natural scenes contain many higher-order forms of statistical structure, and there is good reason to believe they form an extremely non-gaussian distribution that is not at all well captured by orthogonal components¹². Lines and edges, especially curved and fractal-like edges, cannot be characterized by linear pairwise statistics^{6,21} and so a method is needed for evaluating the representation that can

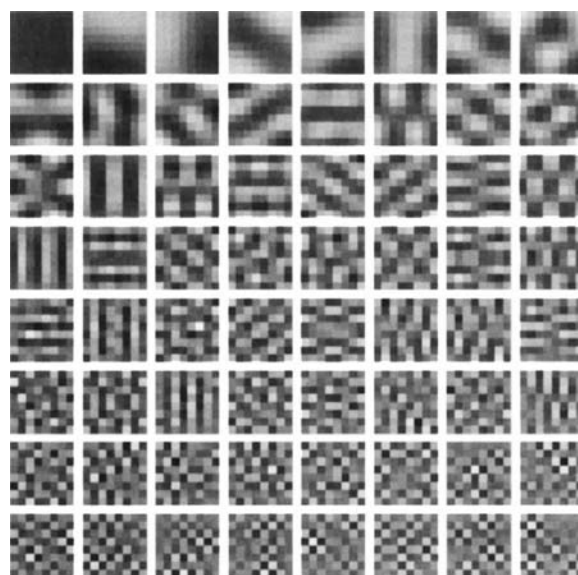


FIG. 1 Principal components calculated on 8×8 image patches extracted from natural scenes by using Sanger's rule¹⁴. The full set of 64 components is shown, ordered by their variance (by columns, then by rows). The oriented structure of the first few principal components does not arise as a result of the oriented structures in natural images, but rather because these functions are composed of a small number of low-frequency components (the lowest spatial frequencies account for the greatest part of the variance in natural scenes⁸). Reconstructions based solely on the first row of functions will merely yield blurry images. Identical-looking components are obtained for images with the same amplitude spectrum as natural images but with randomized phases (that is, $1/f$ noise).

take into account higher-order statistical dependences in the data.

The existence of any statistical dependences among a set of variables may be discerned whenever the joint entropy is less than the sum of individual entropies, $H(a_1, a_2, \dots, a_n) < \sum_i H(a_i)$, otherwise the two quantities will equal. Assuming that we have some way of ensuring that information in the image (joint entropy) is preserved, then a possible strategy for reducing statistical dependences is to lower the individual entropies, $H(a_i)$, as much as possible. In Barlow's terms¹⁹, we seek a minimum-entropy code. We conjecture that natural images have 'sparse structure'—that is, any given image can be represented in terms of a small number of descriptors out of a large set^{8,12}—and so we shall seek a specific form of low-entropy code in which the probability distribution of each coefficient's activity is unimodal and peaked around zero.

The search for a sparse code can be formulated as an optimization problem by constructing the following cost function to be minimized:

$$E = -[\text{preserve information}] - \lambda[\text{sparseness of } a_i] \quad (2)$$

where λ is a positive constant that determines the importance of the second term relative to the first. The first term measures how well the code describes the image, and we choose this to be the mean square of the error between the actual image and the reconstructed image:

$$[\text{preserve information}] = - \sum_{x,y} \left[I(x, y) - \sum_i a_i \phi_i(x, y) \right]^2 \quad (3)$$

The second term assesses the sparseness of the code for a given image by assigning a cost depending on how activity is distributed among the coefficients: those representations in which activity is spread over many coefficients should incur a higher cost than those in which only a few coefficients carry the load. The cost function we have constructed to meet this criterion takes the sum

* Present address: Center for Neuroscience, UC Davis, Davis, California 95616, USA.

of each coefficient's activity passed through a nonlinear function $S(x)$:

$$[\text{sparseness of } a_i] = - \sum_j S\left(\frac{a_i}{\sigma}\right) \quad (4)$$

where σ is a scaling constant. The choices for $S(x)$ that we have experimented with include $-e^{-x^2}$, $\log(1+x^2)$ and $|x|$, and all yield qualitatively similar results (described below). The reasoning behind these choices is that they will favour among activity states with equal variance those with the fewest number of non-zero coefficients. This is illustrated in geometric terms in Fig. 2.

Learning is accomplished by minimizing the total cost functional, E (equation (2)). For each image presentation, E is minimized with respect to the a_i . The ϕ_i then evolve by gradient descent on E averaged over many image presentations. Thus for a given image, the a_i are determined from the equilibrium solution to the differential equation:

$$\dot{a}_i = b_i - \sum_j C_{ij} a_j - \frac{\lambda}{\sigma} S'\left(\frac{a_i}{\sigma}\right) \quad (5)$$

where $b_i = \sum_{x,y} \phi_i(x,y) I(x,y)$ and $C_{ij} = \sum_{x,y} \phi_i(x,y) \phi_j(x,y)$. The learning rule for updating the ϕ is then:

$$\Delta \phi_i(x_m, y_n) = \eta \left\langle a_i \left[I(x_m, y_n) - \hat{I}(x_m, y_n) \right] \right\rangle \quad (6)$$

where $\hat{I}$ is the reconstructed image, $\hat{I}(x_m, y_n) = \sum_i a_i \phi_i(x_m, y_n)$, and η is the learning rate. One can see from inspection of equations (5) and (6) that the dynamics of the a_i , as well as the learning rule for the ϕ_i , have a local network implementation. An intuitive way of understanding the algorithm is that it is seeking a set of ϕ_i for which the a_i can tolerate 'sparsification' with minimum reconstruction error. Importantly, the algorithm allows for the basis functions to be overcomplete (that is, more basis functions than meaningful dimensions in the input) and non-orthogonal⁵, without reducing the degree of sparseness in the representation. This is because the sparseness cost function, S , forces the system to choose, in the case of overlaps, which basis functions are most effective for describing a given structure in the image.

The learning rule (equation (6)) was tested on several artificial datasets containing controlled forms of sparse structure, and the

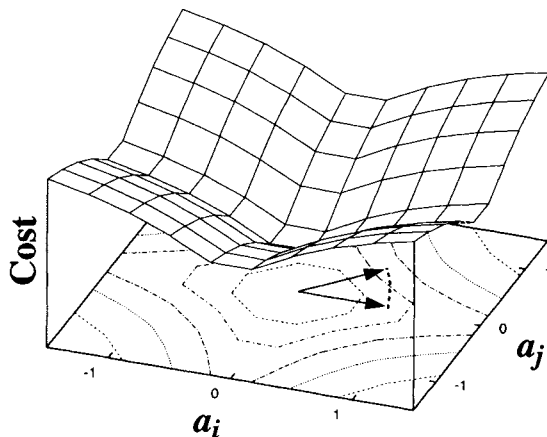


FIG. 2 The cost function for sparseness, plotted as a function of the joint activity of two coefficients, a_i and a_j . In this example, $S(x) = \log(1+x^2)$. An activity vector that points towards a corner, where activity is distributed equally between coefficients, will incur a higher cost than a vector with the same length that lies along one of the axes, where the total activity is loaded onto one coefficient. The gradient tends to 'sparsify' activity by differentially reducing the value of low-activity coefficients more than high-activity coefficients. Alternatively, the sparseness cost function may be interpreted as the negative logarithm of the prior probability of the a_i (ref. 23), assuming statistical independence among the a_i (that is, a factorial distribution), and with the shape of the distribution specified by S (in this case a Cauchy distribution).

results of these tests (Fig. 3) confirm that the algorithm is indeed capable of discovering sparse structure in input data, even when the sparse components are non-orthogonal. The result of training the system on 16×16 image patches extracted from natural scenes is shown in Fig. 4a. The vast majority of basis functions are well localized within each array (with the exception of the low-frequency functions). Moreover, the functions are oriented and selective to different spatial scales. This result should not come as a surprise, because it simply reflects the fact that natural images contain localized, oriented structures with limited phase alignment across spatial frequency⁶. The functions ϕ_i shown are the feedforward weights that, in addition to other terms, contribute to the value of each a_i (refer to term b_i in equation (5)). To establish the correspondence to physiologically measured receptive fields, we mapped out the response of each a_i to spots at every position: the results of this analysis show that the receptive fields are very similar in form to the basis functions (Fig. 4b). The entire set of basis functions forms a complete image code that spans the joint space of spatial position, orientation and scale (Fig. 4c) in a manner similar to wavelet codes, which have previously been shown to form sparse representations of natural images^{8,12,22}. The average spatial-frequency bandwidth is 1.1 octaves (s.d., 0.5) with an average aspect ratio (length/width) of 1.3 (s.d., 0.5), which are characteristics reasonably similar to those of simple-cell receptive fields (~ 1.5 octaves, length/width ~ 2)⁵. The resulting histograms have sparse distributions (Fig. 4d), decreased entropy

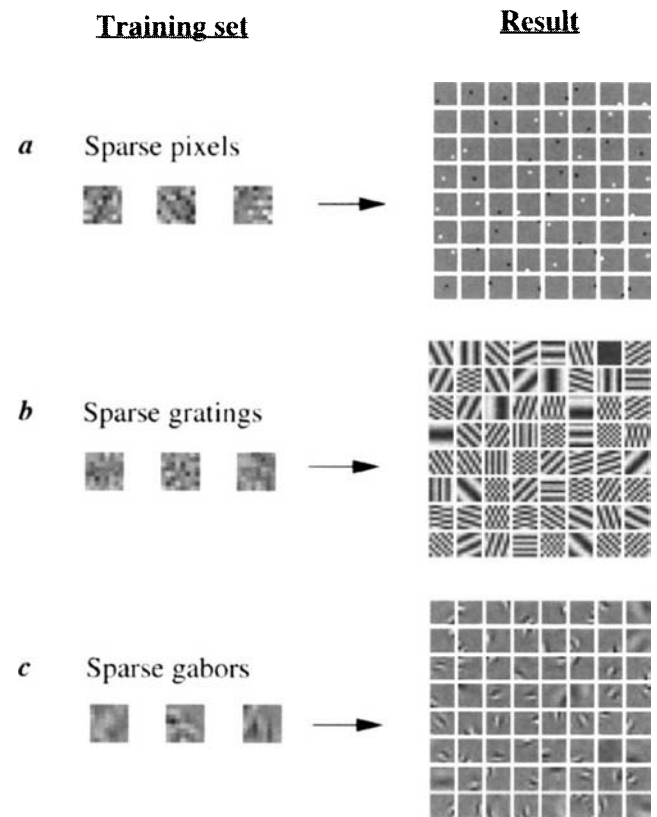
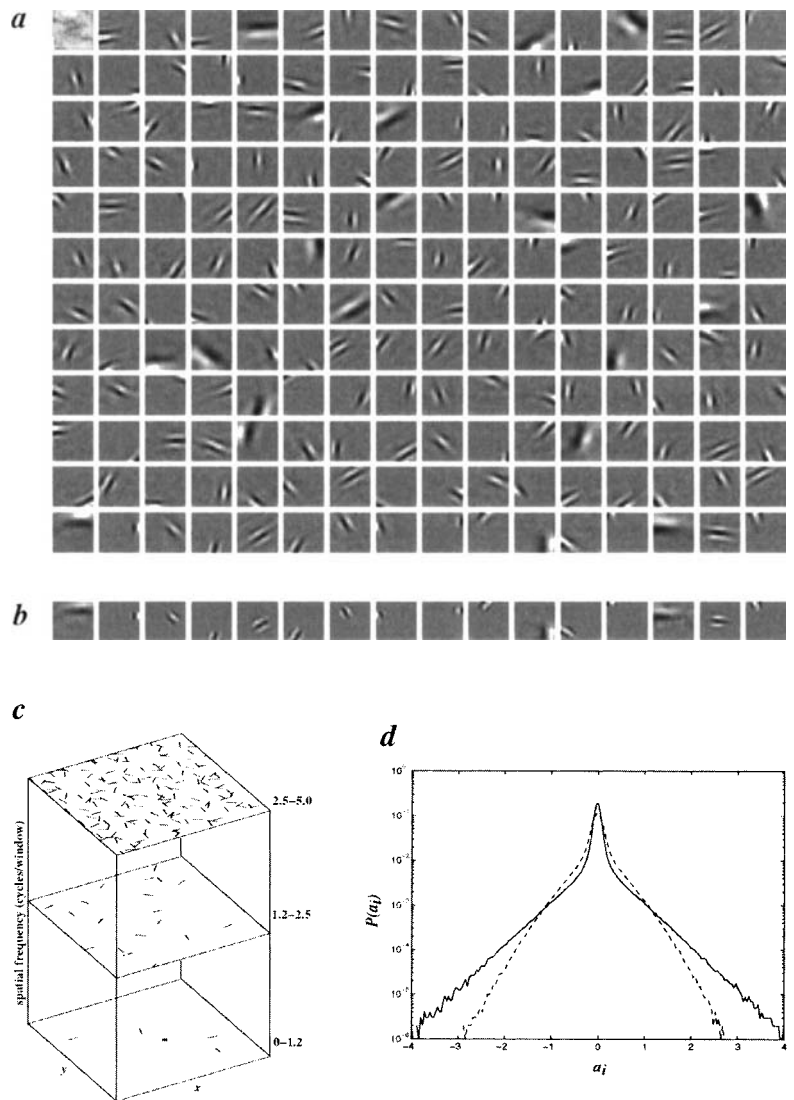


FIG. 3 Test cases. Representative training images are shown at the left and the resulting basis functions that were learned from these examples are shown at the right. In a, images were composed of sparse pixels: each pixel was activated independently according to an exponential distribution, $P(x) = e^{-x}/Z$. In b, images were composed similarly to a, except with gratings instead of pixels (that is, 'sparse pixels' in the Fourier domain). In c, images were composed of sparse, non-orthogonal Gabor functions with the method described by Field¹². In all cases, the basis functions were initialized to random initial conditions. The learned basis functions successfully recover the sparse components from which the images were composed. The form of the sparseness cost function was $S(x) = -e^{-x^2}$, but other choices (see text) yield the same results.

FIG. 4 Results from training a system of 192 basis functions on 16×16 -pixel image patches extracted from natural scenes. The scenes were ten 512×512 images of natural surroundings in the American northwest, preprocessed by filtering with the zero-phase whitening/lowpass filter $R(f) = f e^{-(f/f_0)^4}$, $f_0 = 200$ cycles/picture (see also ref. 9). Whitening counteracts the fact that the mean-square error (or m.s.e.) preferentially weights low frequencies for natural scenes, whereas the attenuation at high spatial-frequencies eliminates artefacts of rectangular sampling. The a_i were computed by the conjugate gradient method, halting when the change in E was less than 1%. The ϕ_i were initialized to random values and were updated every 100 image presentations. The vector length (gain) of each basis function, ϕ_i , was adapted over time so as to maintain equal variance on each coefficient. A stable solution was arrived at after $\sim 4,000$ updates ($\sim 400,000$ image presentations). The parameter λ was set so that $\lambda/\sigma = 0.14$, with σ^2 set to the variance of the images. The form of the sparseness cost function was $S(x) = \log(1 + x^2)$. **a**, The learned basis functions, scaled in magnitude so that each function fills the grey scale, but with zero always represented by the same grey level (black is negative, white is positive). **b**, The receptive fields corresponding to the last row of basis functions in **a**, obtained by mapping with spots (single pixels preprocessed identically with the images). The principal difference may be accounted for by the fact that sparsifying of activity makes units more selective in which aspects of the stimulus they respond to. **c**, The distribution of the learned basis functions in space, orientation and scale. The functions were subdivided into high-, medium- and low-spatial-frequency bands (in octaves), according to the peak frequency in their power spectra, and their spatial location was plotted within the corresponding plane. Orientation preference is denoted by line orientation. **d**, Activity histograms averaged over all coefficients for the learned basis functions (solid line) and for random initial conditions (broken line). In both cases, $\lambda/\sigma = 0.14$, showing that the learned basis functions can accommodate a higher degree of sparsification. Note that even the random basis functions have positive kurtosis due to sparsification. The width of each bin used in calculating the entropy was 0.04.



(4.0 bits compared with 4.6 bits before training), and increased kurtosis (20 compared with 7.0) for a mean-square reconstruction error that is 10% of the image variance.

These results demonstrate that localized, oriented, bandpass receptive fields emerge when only two global objectives are placed on a linear coding of natural images: that information be preserved, and that the representation be sparse. These two objectives alone are sufficient to account for the principal spatial properties of simple-cell receptive fields. A number of unsupervised learning algorithms based on similar principles have been

proposed for finding efficient representations of data^{23–30}, all of which seem to have the potential to arrive at results like these. What remains as a challenge for these algorithms, and also for ours, is to provide an account of other response properties of simple cells (for example, direction selectivity), as well as the complex response properties of neurons at later stages of the visual pathway, which are noted for being highly nonlinear. An important question, then, is whether these higher-order properties can be understood by considering the remaining forms of statistical dependence that exist in natural images. □

Received 10 November 1995; accepted 25 April 1996.

- Hubel, D. H. & Wiesel, T. N. *J. Physiol., Lond.* **195**, 215–244 (1968).
- De Valois, R. L., Albrecht, D. G. & Thorell, L. G. *Vision Res.* **22**, 545–559 (1982).
- Jones, J. P. & Palmer, L. A. *J. Neurophysiol.* **58**, 1233–1258 (1987).
- Parker, A. J. & Hawken, M. J. *J. opt. Soc. Am.* **A5**, 598–605 (1988).
- Daugman, J. G. *Computational Neuroscience* (ed. Schwartz, E.) 403–423 (MIT Press, Cambridge, MA, 1990).
- Field, D. J. in *Wavelets, Fractals, and Fourier Transforms* (eds Farge, M., Hunt, J. & Vascillicos, C.) 151–193 (Oxford Univ. Press, 1993).
- Srinivasan, M. V., Laughlin, S. B. & Dubs, A. *Proc. R. Soc. Lond.* **B216**, 427–459 (1982).
- Field, D. J. *J. opt. Soc. Am.* **A4**, 2379–2394 (1987).
- Atick, J. J. *Network* **3**, 213–251 (1992).
- van Hateren, J. H. *Nature* **360**, 68–70 (1992).
- Ruderman, D. L. *Network* **5**, 517–548 (1994).
- Field, D. J. *Neur. Comput.* **6**, 559–601 (1994).
- Barrow, H. G. in *IEEE First Int. Conf. on Neural Networks* Vol. 4, (eds Caudill, M. & Butler, C.) 115–121 (Institute of Electrical and Electronics Engineers, 1994).
- Sanger, T. D. in *Advances in Neural Information Processing Systems* Vol. 1 (ed. Touretzky, D.) 11–19 (Morgan-Kaufmann, 1989).
- Hancock, P. J. B., Baddeley, R. J. & Smith, L. S. *Network* **3**, 61–72 (1992).
- Law, C. C. & Cooper, L. N. *Proc. natn. Acad. Sci. U.S.A.* **91**, 7797–7801 (1994).
- Fyfe, C. C. & Baddeley, R. *Network* **6**, 333–344 (1995).

- Schmidhuber, J., Eldracher, M. & Foltin, B. *Neur. Comput.* **8**, 773–786 (1996).
- Barlow, H. B. *Neur. Comput.* **1**, 295–311 (1989).
- Linsker, R. *Computer* 105–117 (March, 1988).
- Olshausen, B. A. & Field, D. J. *Network* **7**, 333–339 (1996).
- Daugman, J. G. *IEEE Trans. biomed. Engng.* **36**, 107–114 (1989).
- Harpur, G. F. & Prager, R. W. *Network* **7**, 277–284 (1996).
- Foldiak, P. *Biol. Cybernet.* **64**, 165–170 (1990).
- Zemel, R. S. thesis, Univ. Toronto (1993).
- Intrator, N. *Neur. Comput.* **4**, 98–107 (1992).
- Bell, A. J. & Sejnowski, T. J. *Neur. Comput.* **7**, 1129–1159 (1995).
- Saund, E. *Neur. Comput.* **7**, 51–71 (1995).
- Hinton, G. E., Dayan, P., Frey, B. J. & Neal, R. M. *Science* **268**, 1158–1161 (1995).
- Lu, Z. L., Chubb, C. & Sperling, G. Technical Report MBS 96-15 (Institute for Mathematical Behavioral Sciences, University of California at Irvine, 1996).

ACKNOWLEDGEMENTS. We thank M. Lewicki for helpful discussions at the inception of this work, and C. Lee, C. Brody, G. Harpur, F. Girosi and M. Riesenhuber for useful input. This work was supported by grants from NIMH to both authors. Part of this work was carried out at the Center for Biological and Computational Learning at the Massachusetts Institute of Technology.

CORRESPONDENCE and requests for materials to be addressed to B.A.O. (e-mail: bruno@ai.mit.edu). The program for running the simulation, as well as the images used in training, are available via <http://redwood.psych.cornell.edu/sparsenet/sparsenet.html>.

Primary cortical representation of sounds by the coordination of action-potential timing

R. Christopher deCharms & Michael M. Merzenich

Keck Center for Integrative Neuroscience, 828-HSE, University of California, San Francisco, California 94143, USA

CORTICAL population coding could in principle rely on either the mean rate of neuronal action potentials, or the relative timing of action potentials, or both. When a single sensory stimulus drives many neurons to fire at elevated rates, the spikes of these neurons become tightly synchronized^{1,2}, which could be involved in 'binding' together individual firing-rate feature representations into a unified object percept³. Here we demonstrate that the relative timing of cortical action potentials can signal stimulus features themselves, a function even more basic than feature grouping. Populations of neurons in the primary auditory cortex can coordinate the relative timing of their action potentials such that spikes occur closer together in time during continuous stimuli. In this way cortical neurons can signal stimuli even when their firing rates do not change. Population coding based on relative spike timing can systematically signal stimulus features, it is topographically mapped, and it follows the stimulus time course even where mean firing rate does not.

Although the timing⁴⁻⁸ and correlations⁹⁻¹³ of action potentials have been studied for over three decades, the fundamental logic of cortical population coding is still not understood. The firing rates of the majority of individual neurons in the primary auditory¹⁴⁻¹⁶ somatosensory¹⁷⁻¹⁹ and visual cortices^{20,21} are changed most markedly only at the onset, offset, change or motion of a stimulus, and they typically habituate to nearer their background levels if the stimulus persists, with only a subgroup showing a markedly phasic-tonic profile. Further, it has recently been shown that sensorimotor neurons in the frontal cortex can change their relative timing during behaviour even when their mean firing rates are unchanged²². Are cortical neurons that fire at low rates uninvolved in signalling, or might they be signalling by some other mechanism at the level of the cortical population? Figure 1f,g shows the rate of individual action potentials recorded simultaneously from multiple neurons at each of two locations in the primary auditory cortex of an adult marmoset monkey, which were driven by a pure tone stimulus with the amplitude envelope shown in Fig. 1e, at a frequency appropriate to drive these two locations (4 kHz). Soon after the onset burst, the neurons returned to near their background firing rates, so the firing-rate response to a brief tone pulse (50 ms) was quite similar to that for an ongoing stimulus, although the two stimuli sound dramatically different.

The change in the pattern of relative timing between spikes from these same two locations during these stimuli is shown in Fig. 1a-d. We define the coordination of a population of neurons as the entire pattern of relative action-potential timing within that population, driven by both stimuli and internal dynamics, as this is the signal received by target populations. We define the coordinated rate of a single neuron or group of neurons as the rate of spikes that occur within a fixed time interval from a spike produced by a separate reference neuron or population of neurons. This is computed by making an average peristimulus time histogram of the rate of spikes from one location triggered on spikes from the other (Fig. 1a-d), and measuring the rate in a given bin. This measure reflects both relative timing and relative rate, which both affect postsynaptic targets. Each of the locations shown in Fig. 1 tended to fire more spikes that were close in time to a spike from the other location during the stimuli, even though the overall rate of spikes from each location was not increased, and the neurons were not phase locked to the stimulus (in cycle

histograms, interval distributions, or shift predictors). In both cases, the change in coordinated rate was highly statistically significant ($P < 1 \times 10^{-5}$ for *b*, $P > 3 \times 10^{-4}$ for *c*; permutation test). As a control for changes in relative spike timing or short-term synaptic plasticity that might remain after a stimulus tran-

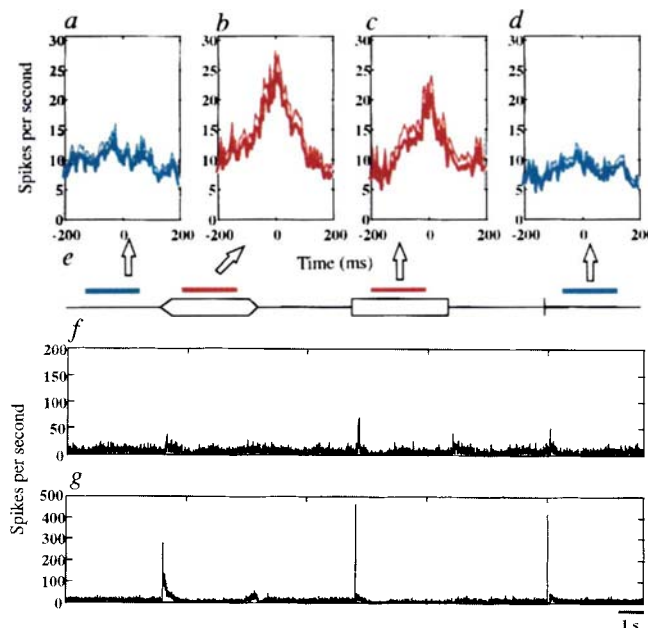


FIG. 1 Changes in mean firing rate and neuronal coordination in the primary auditory cortex during continuous pure-tone stimuli. A 4-kHz pure-tone stimulus with the amplitude envelope shown in e (maximum intensity 70 dB sound pressure level) was presented binaurally during multisite neuronal recording. Histograms of mean neuronal firing rates for two cortical locations are shown in f and g, and average cross-correlations between the two locations are shown in a-d, all constructed from 100 repetitions of the stimulus sequence. a-d, Rate of coordinated discharge (mean, thick line; mean plus standard error, thin line) between the two locations: a, during an initial silent period; b, during the 3-s constant phase of a ramped onset pure tone; c, during the same 3-s constant phase of a rapid onset pure tone; and d, during the silent period following the presentation of a 50-ms pure tone burst. e, Time periods for cross-correlation. The mean half-width at half-height of the central correlation peak in our sample was 11.2 ms (this example is particularly broad). Calculations excluded initial 500 ms after stimulus onsets, and were computed for individual trials by making a spike triggered average of the action-potential rate from the lower cortical location triggered on each action potential from the upper location, using 5 ms bins.

METHODS. We recorded neurons from pairs of locations in the supra-granular primary auditory cortex of 3 marmoset monkeys anaesthetized using Nembutal, with electrode pairs separated by 75–1,000 μ m. We used conventional sharp extracellular microelectrodes, standard electrophysiological amplification with filtering from 300 Hz to 10 kHz, computer storage of spike waveforms sampled at 20 kHz, and computer spike sorting. Data are from neuronal signals that were thresholded and sorted to remove non-neuronal waveforms, giving spike times from 54 pairs of single units with well-differentiated action potentials held throughout (Fig. 2), and 369 pairs of multiunit groups, which consist of the times of all individual action-potential waveforms recorded from each location. Auditory stimuli were generated by a DSP chip and presented binaurally using a closed speaker system. Values of statistical significance of differences between cross-correlograms were derived by calculating the cross-correlation for individual trials using 25-ms bins, which typically captured the central peak, and comparing the values at the zero phase-lag bin between different conditions using the permutation test (a Monte Carlo simulation that does not assume normal distribution)²⁷. They were confirmed using the Mann-Whitney *U* test or joint peristimulus time histogram (JPST) analysis. Mean and s.e. values were derived from all trials, typically 100, and confidence limits were also computed using the point-wise bootstrap method²⁸. We verified our results using two other normalization methods for correlograms: the raw number of coincidences (no normalization), or coincidences normalized by the autocorrelation of the spike trains interpolated to zero phase lag.

sient, we measured the coordinated rate immediately after a brief tone pulse that mimicked the stimulus onset (shown in *d*), and found it to be unchanged. Cortical neurons can thus maintain signals about ongoing stimuli by temporally coordinating the few action potentials present even at low firing rates.

To compare the patterns of coordination found between distinct pairs of individual neurons and within larger neuronal populations, we simultaneously recorded multiple isolated single neurons, multiunit clusters, and local field potentials. The coordinated rate histogram between two single neurons during silence and during a stimulus is shown in Fig. 2*a*. There appeared to be a stimulus-induced change, but there were far too few coordinated spikes at these low firing rates to have sufficient statistical power to measure the effect; this was true of nearly all single unit pairs. A pair of single neurons that did show a statistically significant increase in coordinated rate during the stimulus is shown in Fig. 2*b*, but the effect was still comparatively weak ($P < 0.04$). When we measured the change in average coordinated rate between all of the action potentials taken from two very small groups of neurons (each group containing one of the neurons shown in Fig. 2*b* and one additional cell), we found that the effect was very similar in form but was much more robust ($P < 0.018$). When we added all of the action potentials measured from these same two locations that were above a fixed size, the effect was stronger still ($P < 0.005$). The coordination between one of these single neurons and the local field potential, a measure of the global activity of all surrounding neurons, also showed a similar profile and a significant change ($P < 0.01$). The coordination between nearby pairs of neurons are often similar in form, so their effects can therefore summate and become much more robust in large populations than in single pairs.

Figure 3*a–h* illustrates the frequency tuning of spike time coordination from groups of neurons at two cortical locations that had traditional onset burst tuning to frequencies between 2 and 6 kHz. Groups of neurons from different locations were typically somewhat coordinated even during silence (Fig. 3*a*). Following a brief and frequency-tuned coordinated burst at stimulus onset, the neurons showed an enduring frequency-tuned change in coordinated rate throughout the stimulus (Fig. 3*b–h*). The coordinated rate of action potentials was significantly increased by a 4-kHz stimulus ($P < 0.05$; Fig. 3*e*), but significantly decreased by a 6.35-kHz stimulus ($P < 0.01$; Fig. 3*g*). The coordinated rate can increase even in cases where the overall mean firing rate is decreased (Fig. 3*e*; firing rate less than background, $P < 0.0001$). Where firing rate and relative timing both change, the absolute number of coordinated events reflects both effects.

A particular sound increases the coordinated rate within a restricted and tonotopically mapped sector of the cortical surface (Fig. 3, bottom; the spatial positions of pairs of cortical locations that increased their coordinated rate during the stimulus are connected by red lines, and those that decreased are connected by blue lines). The population response to a 4-kHz tone at 70 dB is shown in Fig. 3*i*, demonstrating that within a discrete subregion of the auditory cortex (containing about 10^6 neurons²³), cells from almost all locations maintain a temporally coordinated pattern of spike timing after a widespread coordinated onset burst; this region may also be decorrelated from surrounding regions. A similar spatial pattern of spike-time coordination is maintained during a stimulus of the same frequency but 30 dB lower intensity (Fig. 3*j*). A stimulus with a nearby but different frequency (2.52 kHz), sampled at fewer points, coordinated a distinct but partly overlapping region (Fig. 3*k*). The centre frequency of coordinated spike tuning versus spatial location for pairs of locations from which we collected full coordination tuning curves is shown in Fig. 3*l*. Coordination tuning forms an organized topographic map across the cortical surface, commensurate with the map for traditional onset burst tuning.

Neuronal coordination could also mimic the time course of the stimulus for locations at which the firing rate did not. Two

locations are shown in Fig. 4*b, c* that did not have any statistically significant change in firing rate in response to a long tonal stimulus, not even transients (because the stimulus rose so slowly; Fig. 4*d*), but still showed a tonic time course of coordinated discharge rate (Fig. 4*a*) that lasted throughout the stimulus. Figure 4*f* shows the sum of all of the firing-rate histograms for this stimulus from the population mapped in Fig. 3. Most of these locations produced a transient discharge, and some also showed smaller sustained changes in firing rate, but the mean of the individual firing rates was unchanged during the constant phase of the stimulus. In contrast, the mean of the coordinated firing-rate profiles from all of the pairs (Fig. 4*e*) showed an increase that lasted throughout the stimulus (Fig. 4*g*). This suggests that a target

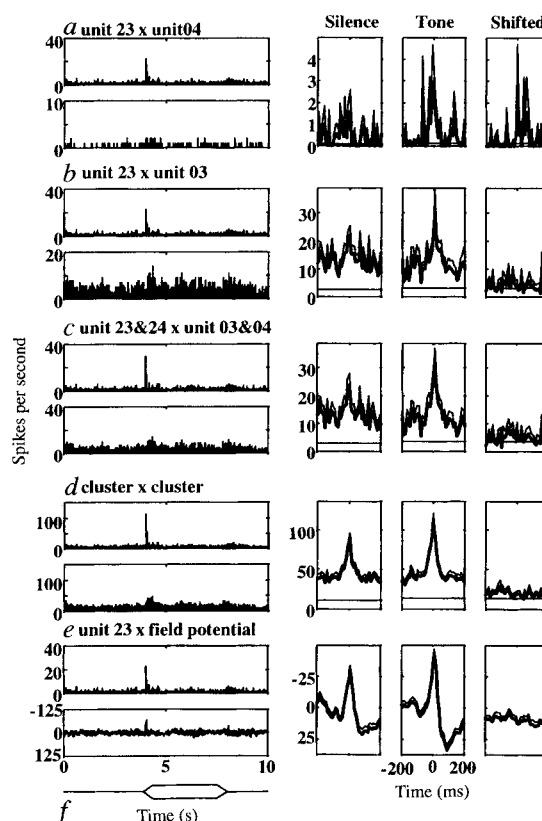


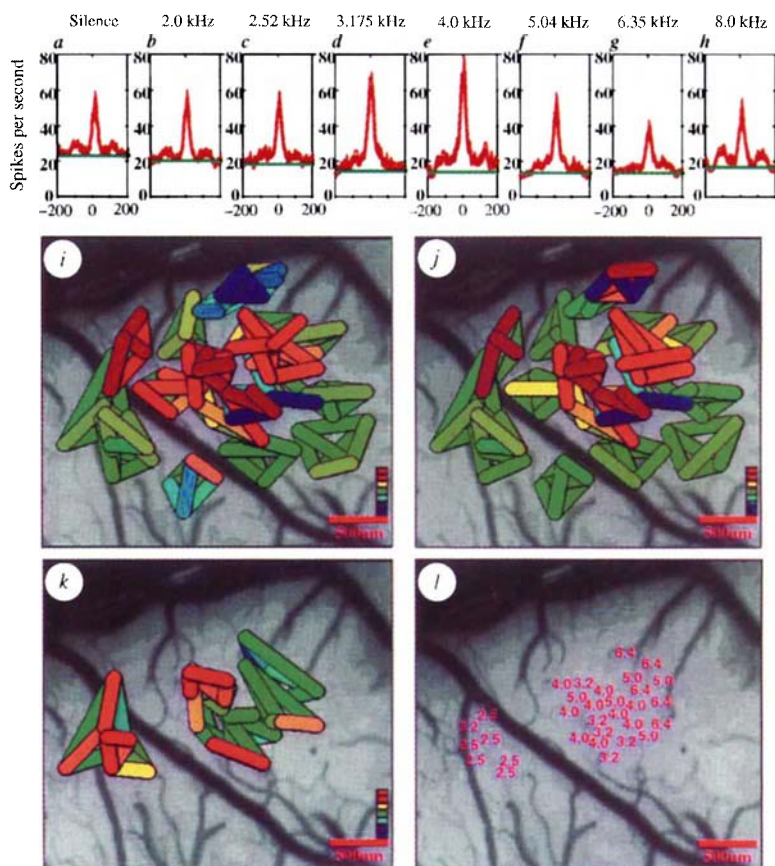
FIG. 2 Neuronal coordination is a population effect. Firing-rate responses from two locations (left: top, reference unit; bottom, target) and the coordination between the two (right) during the silence preceding a stimulus, and during the tonic phase of the stimulus. Far right, the 'shift predictor', a measure of the coordination between two locations that is locked to the timing of the stimulus, computed by making correlations between trial number n for the reference unit and trial $n + 1$ for the target unit. As we are interested in the total coordinated signal to a downstream target, whatever its source, we have not excluded stimulus-locked components using shift predictors of JPST analysis^{29,30}. However, stimulus locking does not contribute to our measured correlation because we chose stimuli during which these neurons do not stimulate lock. None of the shift predictors showed significant stimulus-locked correlation. *a*, Two neurons that appear to increase their cross-correlation during a stimulus, but have too great a variability for the change to be reliably observed. *b*, Two neurons that did show a statistically significant change in cross-correlation due to the stimulus, but had little or no tonic change in firing rate. *c*, Correlation between two groups of well-isolated units, each comprising two neurons and including one of the units shown in *b*, which shows a very similar but more statistically robust effect. *d*, Correlation between two thresholded multiunit groups, which show a similar effect. *e*, Increase in correlation between a single unit and the local field potential recorded from an adjacent electrode (the voltage waveform filtered from 10 to 300 Hz to remove individual action potentials). *f*, Envelope of the presented 4-kHz, 70-dB stimulus. Field potentials are in μV . Mean, thick line; mean plus s.e., thin line.

location which samples across a broad and non-homogeneous population of neurons that are firing on average near background rate can nonetheless extract the time course of an ongoing stimulus by using temporal coincidence information from that population.

Coordinated relative spike timing could in principle be created

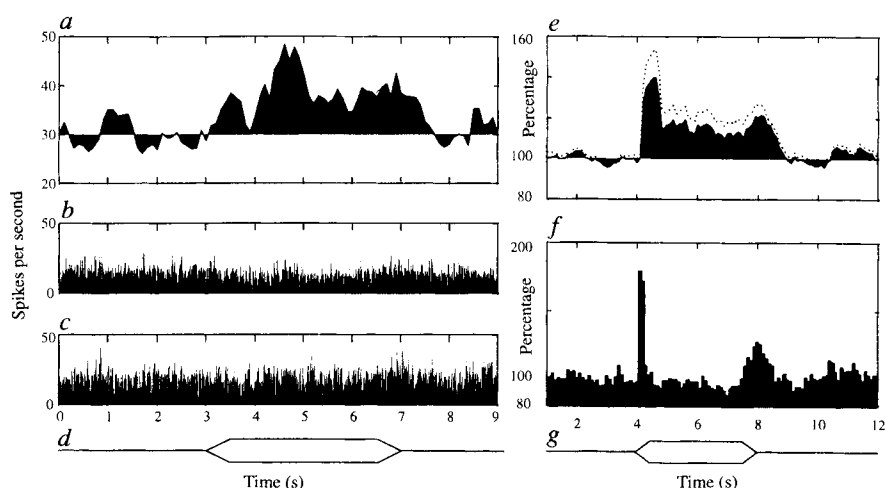
by mechanisms of recurrent neural circuitry, by divergent temporal input to many neurons from a common internal source (possibly the thalamus in this case), or by divergent temporal input from a common external stimulus. All of these sources contribute to the signal received by a postsynaptic cell. Because neurons are temporal integrators, the synchronization and temporal pattern of

FIG. 3 Neuronal coordination is stimulus specific, and can be either increased or decreased relative to background by particular stimuli. *a-h*, Changes in cross-correlation between two locations during silence or 4-s, 70-dB, continuous pure-tone stimuli at different frequencies; mean, thick line; mean plus s.e., thin line. The green lines indicate the mean target-location firing rate during each period. All cross-correlations were computed (as described for Fig. 1) during the tonic phase of the stimulus, excluding onset and offset transient responses. Of 48 pairs of locations for which we constructed full coordination tuning curves using multiple stimuli, 90% showed a stimulus-dependent and frequency-tuned change in coordinated rate ($P < 0.05$ permutation test on central 25-ms bin: of these, 64% tuned increase, 15% tuned increase with surround decrease, 21% tuned decrease), and 10% showed no change. In every case the frequency tuning of ongoing spike coordination was commensurate with that observed for onset bursts, and in some cases was accompanied by changes in firing rate. Of 197 pairs for which we presented only a single stimulus frequency of 4 kHz, which was not the tuned frequency for some neurons, 39.3% showed a significant change in coordinated rate ($P < 0.05$, permutation test). Changes in neuronal coordination also generate an organized topographic map of sound frequency across the cortical surface. We recorded 96 locations in the primary auditory cortex of one monkey in groups of four, each group yielding six measures of cross-correlation, plotted as lines between the corresponding surface locations in *i-k*, with surface vasculature shown for spatial reference. Increases in coordinated rate during the stimulus are shown in red, decreases in blue. *i*, Cortical area that responded with a change in sustained coordination to a 4-kHz tone at 70 dB; *j*, responses of the same locations during presentation of the same tone at 40 dB; *k*, responses of a smaller number of locations during a 2.52-kHz tone at 40 dB; *l*, centre frequency of tonic phase coordinated discharge rate tuning for pairs of locations where a full tuning curve was computed, plotted near the midpoint of the line joining the two locations. Changes in coordination in *i-k* are plotted as the change in coordinated spike rate driven by the stimulus, divided by the background coordinated spike rate,



$(C_{\text{stim}} - C_{\text{back}})/C_{\text{back}}$, with the order of overlaying of lines reflecting the magnitude of this index. The colour bar corresponds to index values of < -1 , -1 , -0.5 , 0.05 , 1 , > 1 from blue to red.

FIG. 4 Cortical sensory neurons can represent a stimulus and its time course through their coordination without any change in mean firing rate. *a*, Time course of the mean coordinated rate in a 500-ms time window ending at the time plotted; *b, c*, histograms (binned to 5 ms) of the evoked firing rates of the neurons at the two cortical locations summed over 100 stimulus presentations; *d*, time course of the stimulus envelope of the 4-kHz, 70-dB tonal stimulus. The coordinated spike rate was computed by calculating the average rate of target-location spikes in a 25-ms bin centred on all reference location spikes during a 500-ms segment of the trial (using the spike-triggered average method as before). Subsequent points in *a* are shifted by 100 ms, leading to $5 \times$ 'oversampling', and shading is centred around the background period mean. The time courses of firing rate and coordinated firing rate were computed for all pairs of locations for this stimulus, and each was normalized to its own background rate. *e*, Mean plus standard error of all coordinated spike-rate time courses, each expressed as a percentage of its own baseline rate. The dotted line indicates the mean plus standard error of all of the curves. *f*, Average of all of the individual firing rates for the same locations, also expressed as a percentage of baseline rate and binned



a 100 ms. The population mean of the coordinated discharge rate remains elevated throughout the time course of a long stimulus (in *g*), while the population mean of individual firing rates marks only the stimulus transients.

inputs arriving at a target cell can, in principle, be used to increase or decrease its output according to basic properties of synaptic integration^{24–26}. At stimulus transients, spikes are necessarily coordinated, so population firing rate and coordinated firing rate reflect essentially the same phenomenon, and these two sides of population signalling presumably work together to provide robust cortical signals in many circumstances. Our results demonstrate that the relative spike timing of cortical action potentials may be a fundamental mechanism for signalling basic object features, one that goes well beyond neuronal grouping.

Received 5 September 1995; accepted 16 April 1996.

- Gray, C. M., König, P., Engel, A. K. & Singer, W. *Nature* **338**, 334–337 (1989).
- Singer, W. & Gray, C. M. *Rev. Neurosci.* **18**, 555–586 (1995).
- von der Malsburg, C. Internal report, Max Planck Institute for Biophysical Chemistry, Göttingen, Germany (1981).
- Adrian, E. D. *Electroenceph. clin. Neurophysiol.* **2**, 377–388 (1950).
- Perkel, D. H., Gerstein, G. L. & Moore, L. *Biophys. J.* **7**, 391–418 (1967).
- Middlebrooks, J. C., Clock, A. E., Xu, L. & Green, D. M. *Science* **264**, 842–844 (1994).
- Abeles, M., Bergman, H., Margalit, E. & Vaadia, E. *J. Neurophysiol.* **70**, 1629–1638 (1993).
- Abeles, M. *Corticonics: Neural Circuits of the Cerebral Cortex* (Cambridge Univ. Press, New York, 1991).
- Dickson, J. W. & Gerstein, G. L. *J. Neurophysiol.* **37**, 1239–1261 (1974).
- Eggermont, J. J. *J. Neurophysiol.* **71**, 246–270 (1994).
- Ts'o, D. Y., Gilbert, C. D. & Wiesel, T. N. *J. Neurosci.* **6**(4), 1160–1170 (1986).
- Ahissar, M. et al. *J. Neurophysiol.* **67**, 203–215 (1992).
- Nicolelis, M. A. L., Baccala, L. A., Lin, R. C. S. & Chapin, J. K. *Science* **268**, 1353–1358 (1995).
- Brugge, J. F. & Merzenich, M. M. *J. Neurophysiol.* **36**, 1138–1158 (1973).
- Creutzfeldt, O., Hellweg, F. C. & Schreiner, C. *Exp. Brain Res.* **39**, 87–104 (1980).
- Pfingst, B. E. & O'Connor, T. A. *J. Neurophysiol.* **45**, 16–34 (1981).
- Mountcastle, V. B., Davies, P. W. & Berman, A. L. *J. Neurophysiol.* **20**, 374–407 (1957).
- Mountcastle, V. B. *J. Neurophysiol.* **20**, 408–434 (1957).

They suggest that a basic cortical signal may be the number of temporally coordinated action potentials in a large neuronal population, which could converge to drive target neurons. Our data can also be understood as a reflection of the temporal patterns of activity that evolve moment by moment in cortical circuitry.

Note added in proof: We have recently observed similar phenomena both in the awake animal and in the primary somatosensory cortex. □

- Mountcastle, V. B., Talbot, W. H., Sakata, H. & Hyvärinen, J. *J. Neurophysiol.* **32**, 452–484 (1969).
- Hubel, D. H. & Wiesel, T. N. *J. Physiol. Lond.* **160**, 106–154 (1962).
- Maunsell, J. H. R. & Gibson, J. R. N. *J. Neurophysiol.* **68**, 1332–1344 (1992).
- Vaadia, E. et al. *Nature* **373**, 515–518 (1995).
- Gabbott, P. L. A. & Stewart, M. G. *Neuroscience* **21**, 833–845 (1987).
- Grannam, E. R., Kleinfeld, D. & Sompolinsky, H. *Neural Comput.* **5**, 550–569 (1993).
- Murthy, V. N. & Fetz, E. E. *Neural Comput.* **6**, 1111–1126 (1994).
- Bernander, O., Koch, C. & Usher, M. *Neural Comput.* **6**, 622–641 (1994).
- Lehmann, E. L. *Nonparametrics: Statistical Methods Based on Ranks* (Holden-Day, San Francisco 1975).
- Efron, B. *The Jackknife, The Bootstrap, and Other Resampling Plans* 75–87 (Society for Industrial and Applied Mathematics, Philadelphia, 1982).
- Perkel, D. H., Gerstein, G. L. & Moore, G. P. *Biophys. J.* **7**, 419–440 (1967).
- Aertsen, A. M., Gerstein, G. L., Habib, M. K. & Palm, G. J. *J. Neurophysiol.* **61**, 900–917 (1989).

ACKNOWLEDGEMENTS. We thank M. Fong, X. Wang, P. Bedenbaugh and D. Buonomano for technical and experimental assistance; E. Ahissar, W. Bialek, D. Ferster, D. Kleinfeld, W. Newsome, W. Singer and others for comments on the manuscript; and K. Parker and J. Stevens for outstanding scientific training. This work was supported by the NIH and by an NSF graduate fellowship to C.d.C.

CORRESPONDENCE and requests for materials should be addressed to M.M.M. (e-mail: merz@phy.ucsf.edu).

Absence of spectrally specific lateral inputs to midget ganglion cells in primate retina

David J. Calkins* & Peter Sterling

Department of Neuroscience, University of Pennsylvania, Philadelphia, Pennsylvania 19104, USA

VISUAL information is conveyed to the brain by retinal ganglion cells. Midget ganglion cells serve fine spatial vision^{1,2} by summing excitation from a receptive field 'centre', receiving input from a single cone in the central retina, with lateral inhibition from a receptive field 'surround', receiving input from many surrounding cones. Midget ganglion cells are also thought to serve colour opponent vision^{1–6} because the centre excitation is from a cone of one spectral type, while the surround inhibition is from cones of the other type^{4,6}. The two major cone types, middle(M)- and long-(L)wavelength sensitive, are equally numerous and randomly distributed in the primate central retina^{7–9}, so a spectrally homogeneous surround requires that the cells mediating lateral interactions (horizontal or amacrine cells) receive selective input from only one cone type. Horizontal cells cannot do this because they receive input indiscriminately from M and L cones^{10–12}. Here we report that the amacrine cells connected to midget ganglion cells are similarly indiscriminate. The absence of spectral specificity in the inhibitory wiring raises doubt about the involvement of midget ganglion cells in colour vision and suggests that colour opponency may instead be conveyed by a different type of ganglion cell.

Amacrine cells extend laterally fine processes (~0.1 µm in diameter) interrupted by varicosities (~0.5–2.0 µm in diameter) containing synaptic vesicles (Fig. 1). The varicosities are the main

sites of excitatory input from bipolar cells and inhibitory output to bipolar and ganglion cells¹³. However, whether the varicosities excited by one midget bipolar cell direct their synapses to the same or to neighbouring midget circuits was unknown (but see refs 5,13). We sought the answer by identifying the amacrine cell processes presynaptic to a midget ganglion cell and its midget bipolar cell, and then tracing them through serial sections to identify their bipolar cell inputs.

A midget ganglion cell dendrite postsynaptic to its midget bipolar cell and surrounded by amacrine cell varicosities (A_{1-9}) is shown in Fig. 1. Two of these varicosities (A_1 and A_8) are postsynaptic to the bipolar cell; A_1 is also presynaptic to the ganglion cell, while A_8 is presynaptic to the bipolar cell. Most varicosities postsynaptic to a particular midget bipolar cell were seen to feed back to the same cell (77%), feed forward to the midget ganglion cell (63%), or do both (53%). Conversely, amacrine cell processes contacting a midget bipolar cell or its ganglion cell (for example, A_2) were usually postsynaptic to the very same bipolar cell (Fig. 1 legend). Thus, for most of the amacrine cell processes associated with midget circuits ('on' as well as 'off'), inputs and outputs are at the same bipolar cell and therefore cannot be spectrally antagonistic.

We next investigated whether there could be cone-selective inputs to other regions of the amacrine cell. We traced and reconstructed three neighbouring amacrine cells (Fig. 2). These cells, although possibly incomplete (and therefore of uncertain type), span and mingle with a patch of 15 midget bipolar cells. A fourth amacrine cell arbor was also reconstructed and spanned a patch of four midget bipolar cells (Fig. 2). The midget bipolar cell array reflects faithfully the cone array, as cone axons do not cross each other¹¹, and neither do midget bipolar cell axons⁸. Thus, we could determine whether inputs to four amacrine cells from a field of 19 midget bipolar cells showed any selectivity. This region also contained terminals from at least five diffuse bipolar cells. These cells collect non-selectively from both M and L cones¹⁴, so it was important to discover whether they also contact the amacrine cells.

Each amacrine cell collected from all the bipolar cells within its reach. Thus, cell 1 received 20 synapses from 10 midget bipolar

* Present address: Max Planck Institute for Brain Research, Deutschordenstrasse 46, D-60528 Frankfurt, Germany.

cells; cell 2 received 8 synapses from 6 midget cells and 5 synapses from diffuse bipolar cells; cell 3 received 8 synapses from 4 midget cells and 2 synapses from diffuse cells; and cell 4 received 8 synapses from 4 midget cells and 3 synapses from diffuse cells. In no case did a midget or diffuse bipolar cell penetrate this thicket of amacrine cell processes without contributing a synapse. That every bipolar cell near these processes, and therefore capable of forming connections, did so makes spectral selectivity for these

amacrine cells quite unlikely. However, we were able to determine the convergence of apparently different spectral types of midget bipolar cell onto these amacrine cells.

Previous studies of our tissue had identified short-wavelength-sensitive (S-cone) 'off' midget bipolar cells^{15,16}. We also partitioned the non-S midget bipolar cells according to whether their terminals were small (~30 ribbons) or large (~50 ribbons)⁸. This distinction apparently divides the population into L and M,

FIG. 1 Dendrite of 'off' midget ganglion cell (MG) receiving two synapses from midget bipolar cell (MB) at synaptic ribbons (r) and surrounded by amacrine cell varicosities (A₁–A₉). Most amacrine cell varicosities postsynaptic to the midget bipolar cell (for example, A₁, A₈) feed back onto the bipolar cell ($77 \pm 5\%$; for example, A₈) or forward onto the midget ganglion cell ($63 \pm 7\%$; for example, A₁); some do both ($53 \pm 9\%$). Many varicosities also contact each other (for example, A₂ → A₃). Conversely, most amacrine cell synapses to the midget bipolar cell ($73 \pm 9\%$) and to its midget ganglion cell ($58 \pm 5\%$) are made by processes postsynaptic to the bipolar cell. Some amacrine cell processes were simply too thin to trace; these may show the same pattern. Thus, for amacrine cells associated with the midget junction, the input and output are at the same site and therefore cannot be spectrally antagonistic.

METHODS. The retina was from an adult, male macaque monkey (*Macaca fascicularis*); preparation of the fovea for electron microscopy is described elsewhere^{8,23}. The material was photographed *en montage* at magnifications of $\times 2,000$, $\times 5,000$ and $\times 15,000$. For 5 'on' and 5 'off' paired midget pathways, the amacrine cell processes contacting the bipolar and ganglion cells were traced to determine whether they were postsynaptic to the same bipolar cell. Conversely, the amacrine processes postsynaptic to each bipolar cell were traced to determine whether they were also presynaptic to the bipolar and/or the ganglion cell. For one 'off' midget ganglion cell, more distant inputs to the presynaptic amacrine cell processes were also traced. These processes gave rise to the three neighbouring amacrine cells shown in Fig. 2, located 620 μm from the centre fovea. A fourth amacrine cell, 45 μm more central, was also reconstructed (Fig. 2, right). The cones providing input to these cells were 1° nasal of the centre fovea^{8,23}.

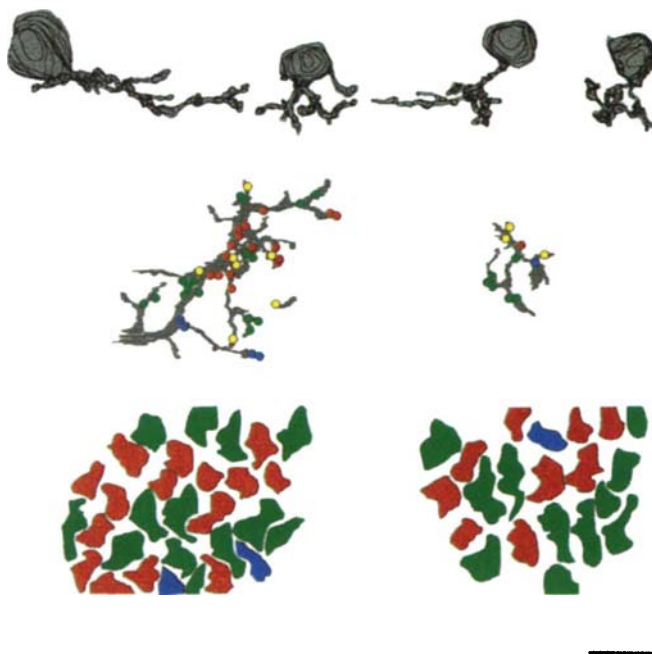
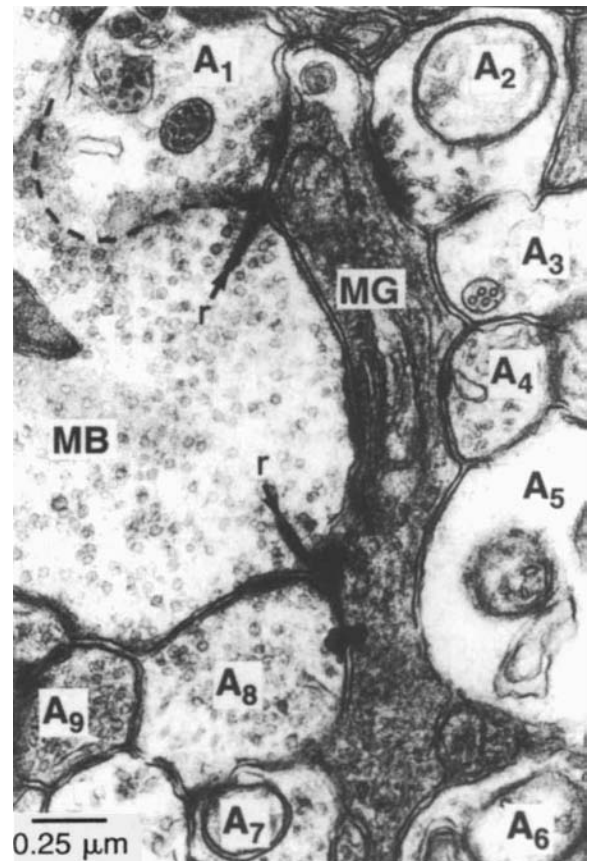
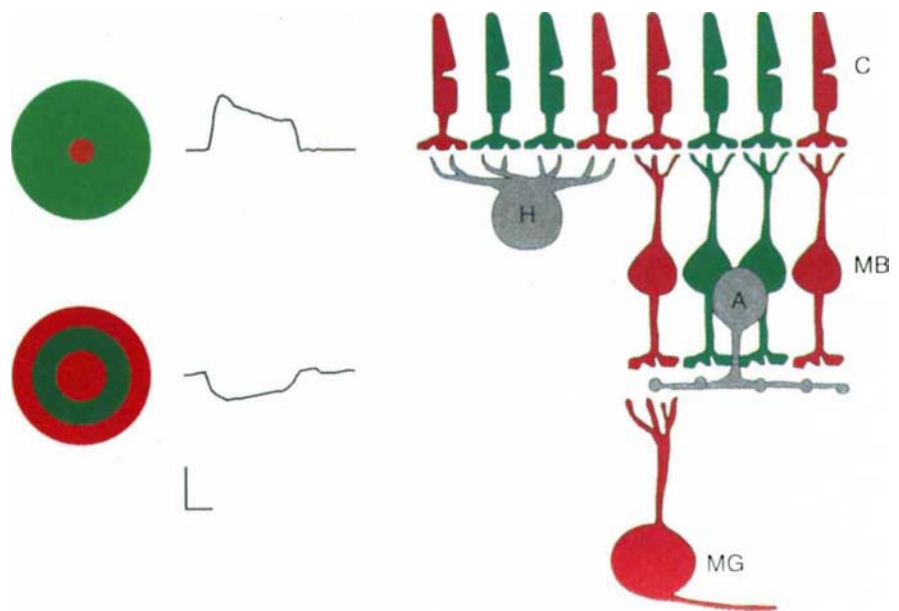


FIG. 2 Top, four amacrine cells in radial view reconstructed from serial sections. The left three cells intertwine *in situ* (middle) but are disentangled here to reveal their individual morphologies. Cell 'types' are uncertain, but two are broad-field and two are narrow-field. The processes are confined to the 'off' region of the inner plexiform layer where they intermingle with the axon terminals of midget and diffuse bipolar cells. Middle, tangential view of the amacrine cell processes intermingling (cell bodies removed). Coloured circles represent synaptic inputs from midget bipolar cells connected to S (blue), M (green) or L (red) cones⁸, or from diffuse bipolar cells (yellow) connected to mixtures of cones (see also ref. 13). As noted in the text, the colour assignment for M and L is arbitrary. The 54 bipolar cell synapses contact the amacrine cell processes non-selectively; there are contributions from every midget and diffuse bipolar cell axon terminal within reach of these processes. Bottom, tangential view of 'off' midget bipolar cell axon terminals in the regions spanned by the four amacrine cells shown above; diffuse bipolar cell terminals (not shown) intermingle with these. A midget ganglion cell centre driven by one midget bipolar cell would tend to have a spectrally broad surround because the cones contributing to the surround by means of horizontal and amacrine cells (Fig. 3) are mixed. Sometimes a bipolar cell is surrounded by a ring of the opposite type (for example, right, red surrounded by green, near bottom of array). The corresponding ganglion cell would show greater spectral antagonism from the surround, but this represents chance and not a fundamental property of the ganglion cell. Scale bar, 10 μm .

FIG. 3 Left, schematic representation of a receptive field of a midget ganglion cell, termed 'type I' (ref. 17). Top, this cell is excited by onset of a red (long wavelength) spot centred on a green (middle wavelength) background. Bottom, cell is inhibited by onset of a green annulus on a red background. Calibration: 10 impulses s^{-1} , 100 ms. Diagram is based on refs 17–22,24,25; traces based on ref. 24. Right, wiring scheme to explain the type I receptive field for midget cells in the central retina or 'fovea'. Centre response is due to excitation of a midget ganglion cell (MG) by, in this example, an L cone (C) through a single⁸ midget bipolar cell (MB). The antagonistic surround formed by only M cones cannot be due to lateral inhibition^{26–28} from the horizontal cell (H) because L and M cone input to horizontal cells is mixed^{10–12}. The surround response has been attributed to selective cone input to amacrine processes^{4,5} (A) by bipolar cells; however, amacrine cells contacting midget bipolar and ganglion cells also collect a mixed cone input (see Fig. 2).



although it does not establish whether the large terminals are L and the small are M, or vice versa. The relevant region of the midget bipolar cell mosaic is classified in this way in Fig. 2 (bottom). The map, when compared with the locations of synapses on the amacrine cells (Fig. 2, middle), shows that both M- and L-cone midget bipolar cells contact each amacrine cell with about equal frequency; the number of contacts between any pair of cells is modest (1–4 synapses). Diffuse bipolar cells also contacted the reconstructed amacrine cells (Fig. 2, yellow symbols). These amacrine cells clearly collect non-selectively from both M and L cones.

Synapses from the three adjacent amacrine cells to any one midget ganglion cell were at most 20% of its total amacrine cell input. Therefore, for one particular 'off' midget ganglion cell, we traced every amacrine cell varicosity providing it a synapse to see whether any would receive selective cone input. Of 33 varicosities, 10 were traced far enough to identify 15 different contributing Off midget bipolar cells, and these were non-selective: 2S, 8M and 5L. At least five neighbouring diffuse bipolar cells also contacted the amacrine cell processes; these bipolar cells collected non-selectively from as many as 10 cones in the same patch of retina (see also ref. 13). Thus, at least 25 cones provide input to these amacrine cells. Our partitioning of M and L midget bipolar cells remains speculative, but the finding that large patches of both

midget and diffuse bipolar cells (with M- and L-cone input) contact amacrine cells non-selectively shows that their input is spectrally mixed.

The anatomical wiring of midget ganglion cells shows both sources of lateral inhibition (horizontal and amacrine cells; Fig. 3, right) to collect from both M and L cones. This is difficult to reconcile with reports^{4,6,17} that surround antagonism is cone selective, however. The random distribution of foveal cones^{7,8} sometimes produces an annulus of one cone type that is either the same or opposite to the cone at its centre (Fig. 2, bottom; Fig. 1e in ref. 8). When most surrounding cones are opposite to the centre cone, then both inhibitory mechanisms cause spectral antagonism, and the receptive field resembles that illustrated in Fig. 3 (originally termed 'type I'; ref. 17). When most surrounding cones are the same type as the centre cone, then both inhibitory mechanisms cause only spatial antagonism, and the receptive field would resemble that originally termed 'type III' (refs 17,18). Commonly, however, the surrounding cones are mixed M and L, so both inhibitory mechanisms cause an intermediate degree of spectral antagonism, which agrees well with other physiological studies^{3,19,20} and efforts to model them²¹. Certainly, the amacrine cell circuitry described here supports the idea that midget ganglion cells, although serving fine-grained spatial vision^{1,2}, may make no special contribution to colour vision²². □

Received 15 December 1995; accepted 24 April 1996.

- Merigan, W. H., Katz, L. M. & Maunsell, J. H. R. *J. Neurosci.* **11**, 994–1001 (1991).
- Schiller, P. H., Logothetis, N. K. & Charles, E. R. *Nature* **343**, 68–70 (1990).
- Derrington, A. M., Krauskopf, J. & Lennie, P. *J. Physiol., Lond.* **357**, 241–265 (1984).
- Reid, R. C. & Shapley, R. M. *Nature* **356**, 716–718 (1992).
- Kolb, H. & Dekorver, L. J. *comp. Neurol.* **303**, 617–636 (1991).
- Lee, B. B. *Vision Res.* **36**, 631–644 (1996).
- Mollon, J. D. & Bowmaker, J. K. *Nature* **360**, 677–679 (1992).
- Calkins, D. J., Schein, S. J., Tsukamoto, Y. & Sterling, P. *Nature* **371**, 70–72 (1994).
- Packer, O. S., Williams, D. R. & Bensinger, D. G. *J. Neurosci.* **16**, 2251–2260 (1996).
- Boycott, B. B., Hopkins, J. M. & Sperling, H. G. *Proc. R. Soc. Lond. B* **229**, 345–379 (1987).
- Wässle, H., Boycott, B. B. & Röhrenbeck, J. *Eur. J. Neurosci.* **1**, 421–435 (1989).
- Dacey, D. M., Lee, B. B., Stafford, D. K., Pokorny, J. & Smith, V. C. *Science* **271**, 656–659 (1996).
- Dowling, J. E. & Boycott, B. B. *Proc. R. Soc. Lond. B* **166**, 80–111 (1966).
- Boycott, B. B. & Wässle, H. *Eur. J. Neurosci.* **3**, 1069–1088 (1991).
- Klug, K., Tsukamoto, Y., Sterling, P. & Schein, S. J. *Soc. Neurosci. Abstr.* **18**, 352.5 (1992).
- Klug, K., Tsukamoto, Y., Sterling, P. & Schein, S. J. *Invest. Ophthalmol. vis. Sci.* (suppl.) **34**, 1398 (1993).
- Wiesel, T. N. & Hubel, D. H. *J. Neurophysiol.* **29**, 1115–1156 (1966).

- Derrington, A. M. & Lennie, P. *J. Physiol., Lond.* **357**, 219–240 (1984).
- Zrenner, E. *Neurophysiological Aspects of Color Vision in Primates* (Springer, Berlin, 1983).
- Kaplan, E., Lee, B. B. & Shapley, R. M. in *Prog. Retin. Res.* **9**, 273–336 (1990).
- Lennie, P., Haake, P. W. & Williams, D. R. in *Computational Models of Visual Processing* (eds Landy, M. S. & Movshon, J. A.) 71–82 (MIT Press, Cambridge, MA, 1991).
- Rodieck, R. W. in *From Pigments to Perception* (eds Valberg, A. & Lee, B. B.) 83–93 (Plenum, New York, 1991).
- Tsukamoto, Y., Masarachia, P., Schein, S. J. & Sterling, P. *Vision Res.* **32**, 1809–1815 (1992).
- DeMonasterio, F. M. *J. Neurophysiol.* **41**, 1418–1434 (1978).
- Croner, L. J. & Kaplan, E. *Vision Res.* **35**, 7–24 (1995).
- Grünert, U. & Wässle, H. *J. comp. Neurol.* **297**, 509–524 (1990).
- Vardi, N. & Sterling, P. *Vision Res.* **34**, 1235–1246 (1994).
- Vardi, N., Kaufman, D. L. & Sterling, P. *Vis. Neurosci.* **11**, 135–142 (1994).

ACKNOWLEDGEMENTS. We thank S. J. Schein for the tissue, and Y. Tsukamoto, P. Masarachia and S. Shrom for preparing and photographing it; H. Wässle for support; S. Fina for help with the illustrations; and B. B. Boycott and R. W. Rodieck for suggestions. This work was supported by research and training grants from the National Eye Institute.

CORRESPONDENCE and requests for materials should be addressed to D.J.C. (e-mail: calkins@mpih-frankfurt.mpg.d400.de).

T-cell-receptor affinity and thymocyte positive selection

S. Munir Alam, Paul J. Travers*, Jay L. Wung, Wade Nasholds*, Stella Redpath, Stephen C. Jameson† & Nicholas R. J. Gascoigne

Department of Immunology, The Scripps Research Institute, 10550 North Torrey Pines Road, La Jolla, California 92037, USA

* Department of Crystallography, Birkbeck College, Malet Street, London WC1E 7HX, UK

† Department of Laboratory Medicine and Pathology, University of Minnesota Medical School, 420 Delaware Street, Minneapolis, Minnesota 55455, USA

DEVELOPMENT of thymocytes involves two distinct outcomes resulting from superficially similar events. Recognition by thymocytes of major histocompatibility complex (MHC) proteins plus peptide leads to their rescue from apoptosis (positive selection), and recognition of antigenic peptide induces cell death (negative selection)¹. Antigen analogues¹⁻³, and sometimes low concentrations of antigenic peptide^{4,5}, induce positive selection; such analogues are often antagonists of mature T-cell clones^{1-3,6}. Various models seek to explain how recognition of different peptide/MHC complexes leads to such different outcomes^{1,7-10}; quantitative models relate response to the affinity, avidity or kinetics of T-cell-antigen receptor (TCR) binding, whereas qualitative models require conformational or spatial changes in the TCR or associated molecules to modulate signal transduction^{7,9}. We have used surface plasmon resonance¹¹ to measure the kinetics of TCR interactions with positively and negatively selecting ligands to distinguish between these models, and find that affinity correlates to the outcome of selection. A 'window' of affinity resulting in positive selection extends over a 1-log range starting threefold below the affinity for negative selection.

The 42.12 TCR recognizes H-2K^b plus an ovalbumin peptide (OVA). Agonist, antagonist and positive-selecting peptides have been defined for this receptor in a system in which the sole variable is the MHC-bound peptide^{2,3,6}. Thus we can isolate the effects of the TCR-MHC/peptide interaction from those of co-receptors or other cell-surface molecules, even though these also contribute *in vivo*¹. Purified complexes of K^b with OVA or VSV (a peptide from

vesicular stomatitis virus) were immobilized on a BIAcore sensor chip, and the binding of soluble (s)TCR was then measured. The 42.12 sTCR, but not an unrelated sTCR, binds to K^b-OVA but not to K^b-VSV (Fig. 1a, b). The binding displays a slow association and a rapid dissociation phase, as previously noted for TCR-MHC/peptide interactions^{12,13}. Titration of 42.12 sTCR binding to K^b-OVA is shown in Fig. 1c. Linear regression analysis of the association phase of these curves (Fig. 1g) gave an association rate (k_{on}) of $3,400 \text{ M}^{-1} \text{ s}^{-1}$ (Fig. 1g) and an estimate of dissociation rate (k_{off}) of 0.03 s^{-1} . Analysis of the dissociation phase of the curves (Fig. 1e) gave a more accurate measurement of k_{off} as 0.023 s^{-1} (Fig. 1f); using this k_{off} value, the association phase gave an estimated k_{on} of $3,412 \text{ M}^{-1} \text{ s}^{-1}$. From these rates we calculate the equilibrium dissociation constant (K_d) for the interaction of 42.12 TCR with K^b-OVA as $6.83 \times 10^{-6} \text{ M}$.

To verify our data, we reversed the ligand/analyte orientation by immobilizing the sTCR. Injections of K^b-OVA and K^b-VSV showed specific binding only for K^b-OVA (Fig. 2a) over the range 4–11 μM (Fig. 2c). An analysis similar to that outlined for the data in Fig. 1 determined k_{on} as $3,116 \text{ M}^{-1} \text{ s}^{-1}$ and k_{off} as 0.017 s^{-1} (Fig. 2g). Binding the sTCR to the sensor chip through immobilized anti-C β antibody (Fig. 2e, f) gave similar results ($k_{on} = 3,456 \text{ M}^{-1} \text{ s}^{-1}$, $k_{off} = 0.019 \text{ s}^{-1}$; Fig. 2h). The calculated K_d values for the interaction between TCR and K^b-OVA from these two measurements were very close: $5.54 \times 10^{-6} \text{ M}$ versus $5.64 \times 10^{-6} \text{ M}$. These are also close to the K_d obtained when K^b-OVA was immobilized.

Similar off-rates were found previously for a class I alloreactive (0.026 s^{-1})¹² and a class II-restricted TCR ($0.06\text{--}0.09 \text{ s}^{-1}$)¹³. For the class II system, the on-rate was slightly slower ($900\text{--}1,700 \text{ M}^{-1} \text{ s}^{-1}$)¹³ than we observe, while for the class I system it was much faster ($2.1 \times 10^5 \text{ M}^{-1} \text{ s}^{-1}$)¹², resulting in much higher affinities. Values of $1 \times 10^{-7} \text{ M}$ to $5 \times 10^{-8} \text{ M}$ were reported, although some agonists gave affinities as low as $3 \times 10^{-4} \text{ M}$ ^{12,14,15}. The K_d of another K^b-OVA-specific TCR was estimated as $0.67 \mu\text{M}$, a 10-fold higher affinity than we found¹⁵. This may not be significant, as similar discrepancies have been found between measurements made in BIAcore and with whole cells^{12,14,15}.

To determine whether the outcome of recognition of different MHC/peptide complexes correlates with differences in their affinity for the TCR, we repeated our measurements using a series of positive-selecting/antagonist peptides (Fig. 3 and Table 1). The peptides V-OVA and R4 are strict antagonists, whereas E1 is a stronger antagonist and a weak agonist^{2,3}. These peptides

TABLE 1 Kinetics for TCR 42.12 binding to various H-2K^b complexes

K ^b peptide	k_{on} ($\text{M}^{-1} \text{ s}^{-1}$) ($\pm$ s.e.m.)	k_{off} (s^{-1}) ($\pm$ s.e.m.)	Half-life (s)	K_d (μM) off/on	Thymocyte selection*	Mature T-cell activation*
OVA (SIINFEKL)	3135 ± 75 †	0.020 ± 0.001	24.5	6.5	negative	antigen
E1 (EIINFEKL)	3023 ± 371	0.068 ± 0.004 †	7.3	22.6	positive	weak agonist/strong antagonist
V-OVA (RGYNYEKL)	1297 ± 261 †	0.039 ± 0.005 †	12.9	29.8	weak positive	antagonist
R4 (SIIRFEKL)	2565 ± 39 †	0.146 ± 0.012 †	3.4	57.1	positive	antagonist
K4 (SIIKFEKL)	(552)	(>0.2)	(<2.5)	(>362)	none	none
VSV (RGVYQGL)		(>0.6)	(<0.8)		none	none

Values of k_{on} and k_{off} were determined for a series of K^b-peptide complexes using curve fitting to a simple two-component model of interaction ($AB = A + B$) for a titration of the solution-phase ligand, as for Figs 1 and 2. The experiments were performed with either the K^b complexes or the TCR immobilized. Mean values ($\pm$ s.e.m.) were calculated after pooling measurements made by both methods. Values in brackets are approximate owing to the very slow association and fast dissociation rates for these complexes. Individual k_{on} values ($\text{M}^{-1} \text{ s}^{-1}$) determined with the K^b-peptide complex immobilized were as follows: K^b-OVA, 3,412, 2,770; K^b-E1, 3,285; K^b-V-OVA, 1,890; K^b-R4, 2,510; K^b-K4, ~552. Determinations with the TCR immobilized were as follows: K^b-OVA, 3,116, 2,920, 3,456; K^b-E1, 2,760; K^b-V-OVA, 900, 1,100; K^b-R4, 2,530, 2,655. Despite several attempts, we were unable to estimate the on-rate for the K^b-VSV complex, and that for K^b-K4 could only be estimated in one experiment. This measurement is too slow to be considered accurate. Individual k_{off} values (s^{-1}) from measurements with K^b-peptide immobilized were: K^b-OVA, 0.023, 0.025; K^b-E1, 0.070; K^b-V-OVA, 0.051; K^b-R4, 0.120; K^b-K4, ~0.20; K^b-VSV, ~0.85. Values determined from experiments with TCR immobilized were: K^b-OVA, 0.018, 0.017, 0.019; K^b-E1, 0.060, 0.075; K^b-V-OVA, 0.035, 0.030; K^b-R4, 0.150, 0.169; K^b-K4, ~0.50; K^b-VSV, ~0.55. Of the measurements with TCR immobilized, two of the OVA values, the E1 measurement, and one each of the V-OVA and R4 measurements were obtained using TCR bound through immobilized anti-C β .

* Information from refs 2, 3, 6.

† Values of rates significantly different ($P < 0.05$, Student's *t*-test) from K^b-OVA binding. On-rates: OVA/V-OVA, $P = 0.001$; OVA/R4, $P = 0.017$; E1/V-OVA, $P = 0.029$; V-OVA/R4, $P = 0.014$. OVA/E1 and E1/R4 were non-significant. Off-rates: OVA/E1, $P = 0.001$; OVA/V-OVA, $P = 0.012$; OVA/R4, $P = 0.001$; E1/V-OVA, $P = 0.018$; E1/R4, $P = 0.006$; V-OVA/R4, $P = 0.002$.

differ in their ability to induce positive selection; E1 is a strong selector, R4 is less strong and V-OVA weaker still². We also tested a non-stimulatory peptide, K4, which neither antagonizes nor positively selects the 42.12 receptor. The binding of 42.12 sTCR showed little difference between K^b-E1, K^b-R4 and K^b-OVA in the association phase (Fig. 3a), but K^b-V-OVA was 2.4-fold slower, and K^b-K4 more than 6-fold slower (Fig. 3c and Table 1). The positive-selecting peptide complexes dissociated 2–7-fold faster than the K^b-OVA complex (Fig. 3b, d and Table 1). However, K^b-V-OVA, a weak positive selector, was

the closest to K^b-OVA in k_{off} (significantly different, $P = 0.012$). The non-stimulatory complexes K^b-K4 and K^b-VSV dissociate faster than either positive or negative selectors. By using the on-rate for K^b-K4 also for K^b-VSV (clearly an overestimate), we calculate the equilibrium dissociation constants (K_d) to be in the order of OVA (6.5 μM) < E1 (22.6 μM) < V-OVA (29.8 μM) < R4 (57.1 μM) $\ll$ K4 (> 362 μM) $\ll$ VSV (> 1,000 μM). Thus, all of the positive selecting MHC/peptide complexes have a lower affinity (3–9-fold) than K^b-OVA. We also estimated the K_d for selected complexes by Scatchard analysis of equilibrium binding

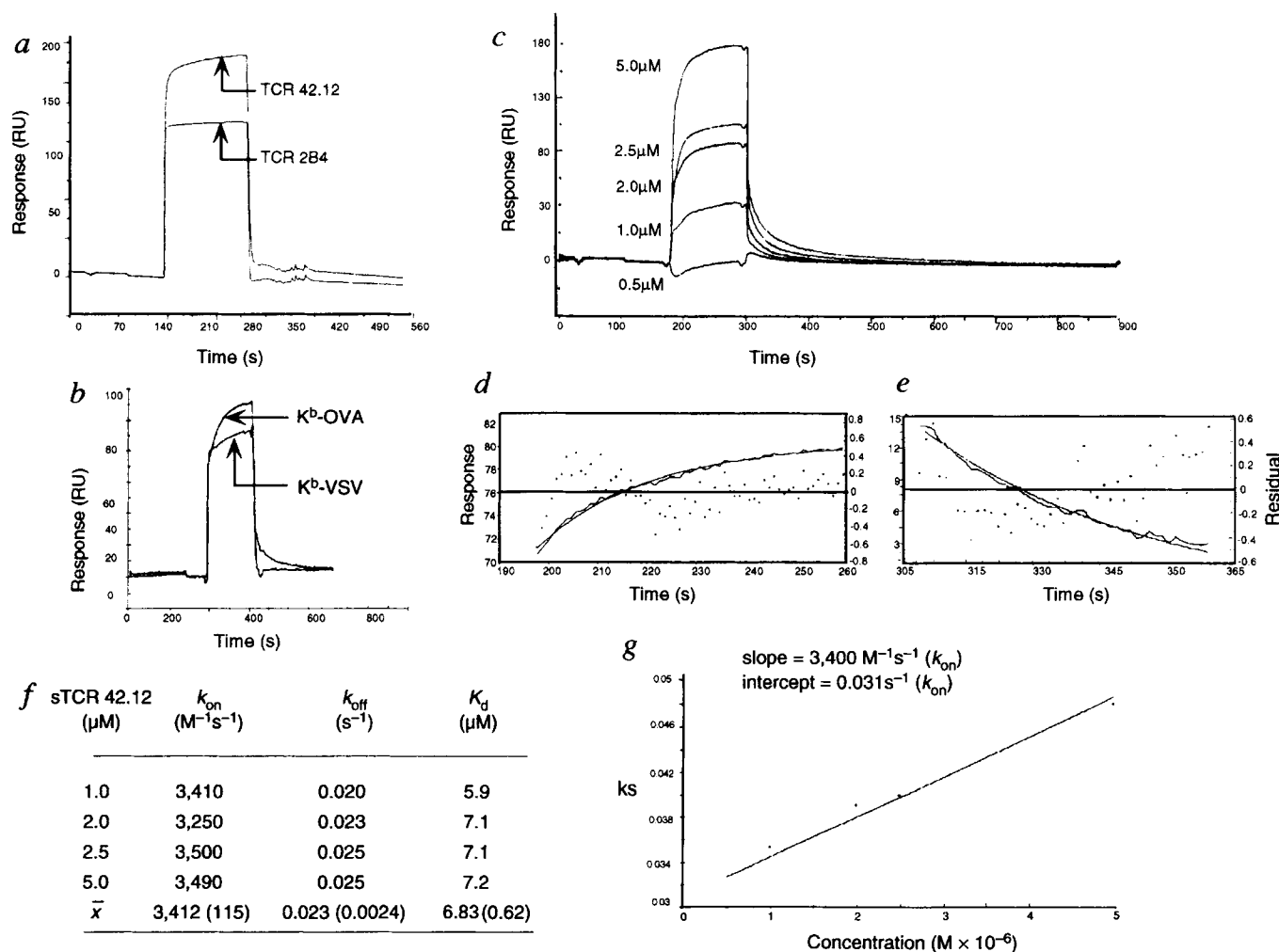


FIG. 1 Surface plasmon resonance measurement of the interaction of sTCR with immobilized K^b-peptide complexes. **a**, TCR 42.12, but not control sTCR 2B4 (ref. 13), binds to the antigenic K^b-OVA complex (2B4 sTCR bound the bacterial superantigen SED; data not shown). **b**, TCR 42.12 binds poorly to K^b-VSV compared to K^b-OVA. **c**, Kinetic analysis of concentration-dependent binding of sTCR to K^b-OVA. These curves gave a good fit to a two-component binding interaction ($A + B = AB$); **d** and **e** are representative examples of the curve fitting and the residual plots of the association ($\chi^2 = 0.05$) and dissociation ($\chi^2 = 0.065$) phases, respectively. **f**, The k_{on} and k_{off} values were calculated using the $A + B = AB$ model fit for each of the sTCR concentrations. The equilibrium dissociation constant, K_d , was calculated from $k_{\text{off}}/k_{\text{on}}$. Numbers in brackets represent $\pm$ s.d. **g**, Linear regression analysis of the curves in **c**, as a k_s concentration (of TCR) plot (see ref. 11; k_s is the slope of dR/dt versus R , where R is the response (in resonance units, RU). The slope of this curve gives an apparent k_{on} of $3,400 \text{ M}^{-1}\text{s}^{-1}$. An apparent k_{off} of 0.03 s^{-1} was estimated from the intercept with the y-axis.

METHODS. The 42.12 T cell recognizes an ovalbumin (OVA) peptide in association with H-2K^b. Positive/negative-selecting and agonist/antagonist peptides were determined for its TCR^{2,3,6}, consisting of AV2S7J33 and

BV5S2D2J2S6 (ref. 20; see refs 21,22 for nomenclature). A control AV11S2J56, BV3S1A1D2J2S5 sTCR was made from the cell line 2B4 (ref. 13). The soluble forms of the TCR proteins were expressed from truncated genes in *Drosophila* S2/M3 cells (expression vector pRM Ha-3)²³, or from the yeast *Pichia pastoris*, using the vectors pPIC9 for α -chain and pPIC9K for β -chain (Invitrogen, San Diego)²⁴. The truncation involved deletion of the transmembrane region (in the yeast system the cytoplasmic tail was also deleted)^{25,26} with the addition of six histidine residues at the C terminus. Purification was on a Ni-NTA column followed by gel-filtration on a Superdex HR-200 column. The K^b-peptide complexes were prepared essentially as described²⁷. Correctly refolded K^b-peptide complexes were purified on gel-filtration high performance liquid chromatography. Before use, the proteins were run through a Superdex HR200 column to remove aggregates. K^b-peptide complexes (1,200–1,500 RU) were amine-coupled to a CM5 sensorchip in a BIAcore 2000 Biosensor (Pharmacia) in 10 mM acetate buffer, pH 4.0. The immobilized protein was 15–25% active as detected by conformation-specific mAb Y3 (ref. 28). Binding was monitored at 25 °C; flow rate, $10 \mu\text{l min}^{-1}$; buffer, PBS (150 mM NaCl, 0.1% surfactant P20, pH 7.4). Analyses used a nonlinear fit method with BIAevaluation 2.0 software (Pharmacia).

(Fig. 3e, f). This gave slightly higher K_d values, but with the same hierarchy: OVA (18.8 μM) < E1 (34.9 μM) < V-OVA (54.9 μM). Similar differences (2–3-fold) between the K_d values determined from kinetic measurements and those determined from equilibrium measurements have been found previously¹². We estimate that a K_d of $\sim 10 \mu\text{M}$ and lower (the K_d for K^b -OVA being 6.5 μM) would result in negative selection of this TCR. Positive selection is evident at a 3-fold lower affinity, and extends over at least a 2.5-fold (20–50 μM) range. The non-selecting K^b -K4 complex has a K_d of $>362 \mu\text{M}$, providing a lower boundary for the range of affinities capable of supporting positive selection of this receptor. In this system, therefore, a 50-fold difference in affinity encompasses the range from no selection to negative selection.

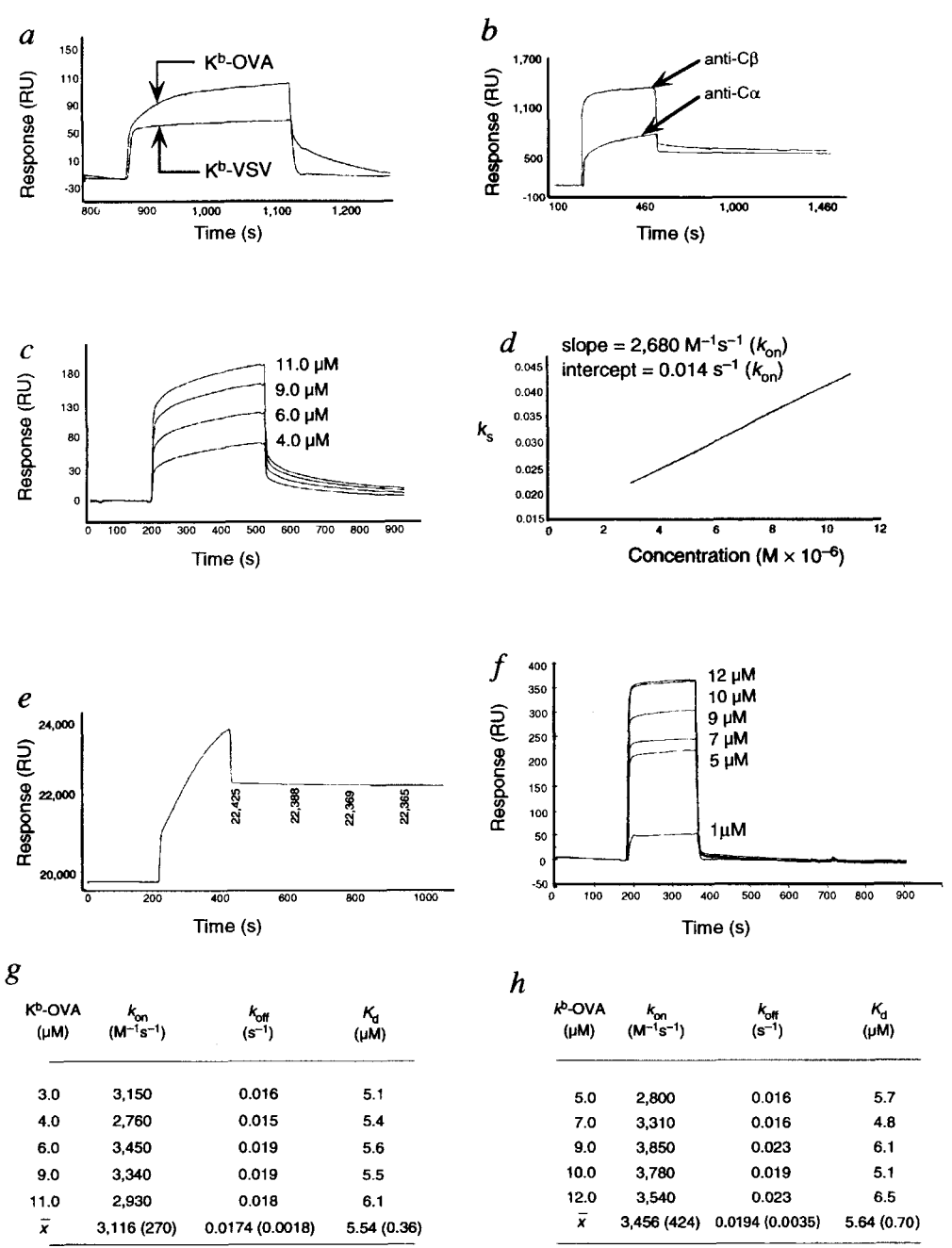
We found a general correlation between affinity of the TCR-peptide/MHC interaction and the outcome in thymic selection and T-cell activation. In a class I allorecognition system, affinity correlated to response of mature T-cell clones for a range of agonists^{14,16}, with one discrepancy¹⁶. Of the affinity-based models

for signal transduction, the most attractive are the kinetic proof-reading⁸ or discrimination¹⁰ models. In these the complexity of the signalling process⁸, or the ratio of complete (positive) to incomplete (negative) signals¹⁰, allows small differences in affinity to yield large differences in outcome^{8–10}; this matches the experimental results. Differences in off-rates are the distinguishing factors between ligands in these models, and such a relationship has been suggested in a class-II-restricted system¹³. However, we did not find a strict correlation between off-rate and responsiveness. Data suggesting this correlation derive from studies testing single-residue changes in the peptide ligand¹³. Had we not included V-OVA, which differs from OVA in four of eight residues, we would also have concluded that the off-rate was the key discriminating factor. Other experimental systems may therefore be biased to that conclusion by the choice of peptides.

It has been proposed that the total number of receptors ligated during the interaction between the T-cell and the antigen-presenting cell (APC) determines the outcome¹⁷. Again, off-rates are

Fig. 2 Interaction of K^b -peptide complexes to immobilized TCR. a, K^b -OVA (5 μM) bound significantly to immobilized sTCR but K^b -VSV bound very poorly. b, Immobilized TCR bound an equivalent mass of anti-C β (H57–597) and anti-C α (H28–710) mAbs (see refs 25,26). c, Titration of K^b -OVA binding to immobilized TCR. The bulk refractive index was subtracted from each of the curves to produce the true kinetic curves¹², which fit to the AB = A + B model (association, $\chi^2 = 0.26$; dissociation, $\chi^2 = 0.8$). d, Linear regression analysis of a k_s versus concentration plot. e, TCR was immobilized by binding to an anti-C β mAb surface. The antibody-bound sTCR was stable with negligible baseline drift. The RU of bound TCR remaining at different time intervals are indicated. f, Concentration-dependent binding of K^b -OVA to the TCR/anti-C β surface. The curves gave a good fit to the AB = A + B model by curve fitting and residual plots (not shown; association, $\chi^2 = 0.065$; dissociation, $\chi^2 = 0.08$). g, h, The apparent k_{on} and k_{off} were determined for each of the analyte concentrations shown in c and f for the direct TCR immobilization (g) and indirect immobilization (h).

METHODS. For a–d, g, 1,500–2,000 RU of sTCR protein was directly immobilized (20–25% active as judged by TCR-specific mAbs; see Fig. 1 legend). The K^b -OVA or K^b -VSV were injected as analyte. For e, f, h, the sTCR was captured by anti-C β mAb (immobilized using acetate buffer, pH 4.5); TCR immobilized in this was 50–90% active. After binding to the anti-C β mAb, sufficient washout time was allowed to exclude any contribution from baseline drift (e). This provided a significantly improved TCR surface to study the nature of binding of sTCR to K^b -peptide complexes. Direct immobilization of the TCR gave rise to artefacts suggesting a more complex interaction than the simple two-component AB = A + B model. Similar problems¹² or complete inactivity of the TCR²⁹ have been seen previously on direct immobilization of TCR. These were reduced by subtraction of binding to a blank chip, allowing analysis of the data by AB = A + B. Immobilization through the anti-C β mAb gave much cleaner results. The possibility of a true complex interaction is excluded because binding to both the K^b -OVA surface (Fig. 1) and the TCR/anti-C β surface fitted to the simple model without manipulation.



critical in this model¹⁰. Because V-OVA has an off-rate intermediate between those of OVA and E1, it should be capable of demonstrating agonist activity. However, V-OVA is unable to act as an agonist at any concentration up to 1,000-fold greater than that required to show maximal activity for OVA or E1 (refs 2,6).

In the interaction between a T cell and an APC, the TCR, co-receptors, specific MHC/peptide complexes and cell adhesion molecules colocalize at points of contact between the T cell and the APC¹⁸. Such clustering results in high concentrations of these molecules. In the case of the DC2:LFA-3 interaction, which has a K_d comparable to that of the 42.12 TCR, clustering results in effective ligand concentrations that are higher than the K_d for the interaction¹⁹. We might thus expect T-cell activation to be related to affinity or avidity rather than off-rate. Our data suggest that avidity is the relevant parameter, because the K^b -V-OVA complex has a higher affinity for the TCR than does K^b -R4, but is a less efficient positive selector. However, V-OVA binds less well to H-2K^b than R4 (ref. 2). This difference may decrease the avidity,

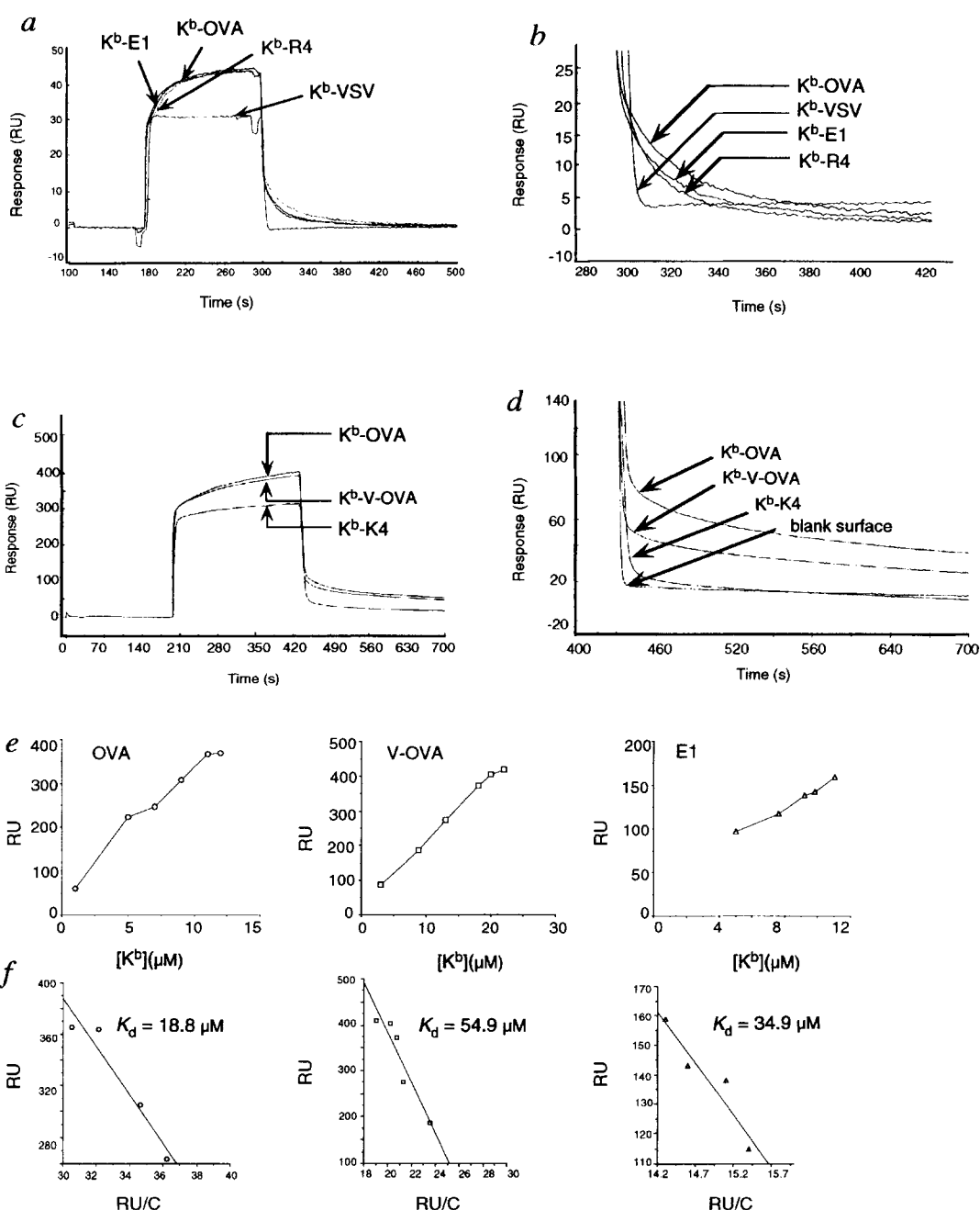
or efficacy², of the K^b -V-OVA interaction sufficiently to account for the lower responsiveness to this complex.

Our data argue against models that require ligand-induced conformational changes in either the TCR or MHC molecules. Ligands that do induce conformational changes should show slower on- and/or off-rates than those that do not; the slow association rate for TCR and its ligands has been suggested to be consistent with such a model¹³. The rates for the antagonist peptide complexes are actually slower than those for the agonist complex. In particular, the K^b -V-OVA complex is a weak positive selector, and would be expected not to induce any necessary conformational change in the TCR, yet its on-rate is slower than for the K^b -OVA complex, and its off-rate is slower than for K^b -E1, a strong positive selector. We cannot exclude a more broadly defined conformational model in which the ligands have a differential effect on TCR dimerization or co-receptor binding⁹.

The affinity of the TCR is one part of a complex interaction between cell surfaces in which the densities of the receptor,

FIG. 3 Interaction of TCR with antagonist/positive-selecting K^b -peptide complexes. **a**, Binding of sTCR 42.12 to antagonist/positive-selecting ligands K^b -E1 and K^b -R4 compared to antigenic K^b -OVA. All three complexes bound with a higher signal than K^b -VSV, and showed very similar association curves. The dissociation phase of the interaction is shown magnified in **b**. Antagonist/positive-selecting complexes dissociated faster than the antigenic complex. **c**, Binding of K^b -OVA compared to the weak positive-selector K^b -V-OVA and non-selecting K^b -K4. The association curve for K^b -V-OVA is slow compared to K^b -OVA. **d**, The dissociation curve in **c** is magnified showing that the non-selecting complex dissociates much faster than either of the stimulatory complexes, and that the positive-selecting complex dissociates faster than the negative-selecting complex. **e**, Equilibrium binding data (RU versus C plot) obtained by capturing sTCR proteins on an anti-C β immobilized surface followed by injections of OVA, V-OVA and E1. **f**, Scatchard plots of the data in **e**; the equilibrium dissociation constant, K_d , was calculated from the slope of the fitted line.

METHODS. As for Figs 1 and 2d, **e**, **g**. To allow visual comparison between the kinetics for the different complexes, curves reaching approximately the same maximal RU as that obtained with the K^b -OVA complex are shown. The curves in **a** and **b** were derived by injecting 2 μ M K^b -OVA and 5 μ M of the other complexes. For **c** and **d**, K^b -V-OVA was injected at 18 μ M, the others at 10 μ M. For equilibrium binding experiments, each of the MHC complexes was injected over a TCR/anti-C β surface with the same level of TCR bound ($\sim$ 800 RU). This was ensured by regenerating the surface after each MHC-TCR interaction and injecting the same amount of TCR. The quality of the different K^b complexes was equivalent, with almost identical binding to immobilized conformation-specific anti- K^b (Y3) mAb.



co-receptors, adhesion molecules and potential selecting ligands play a part. Our results show that very small differences in the affinity of one of these interactions, between the receptor and the selecting ligand, are sufficient to alter the fate of the developing thymocyte. □

Received 14 March; accepted 19 April 1996.

1. Jameson, S. C., Hogquist, K. A. & Bevan, M. J. *A. Rev. Immun.* **13**, 93–126 (1995).
2. Hogquist, K. A. et al. *Cell* **76**, 17–27 (1994).
3. Jameson, S. C., Hogquist, K. A. & Bevan, M. J. *Nature* **369**, 750–752 (1994).
4. Ashton-Rickardt, P. G. et al. *Cell* **76**, 651–663 (1994).
5. Sebzdka, E. et al. *Science* **263**, 1615–1618 (1994).
6. Jameson, S. C., Carbone, F. R. & Bevan, M. J. *J. exp. Med.* **177**, 1541–1550 (1993).
7. Janeway, C. A. Jr *Immun. Today* **16**, 223–225 (1995).
8. McKeithan, T. W. *Proc. natn. Acad. Sci. U.S.A.* **92**, 5042–5046 (1995).
9. Germain, R. N., Levine, E. H. & Madrenas, J. *Immunologist* **3/4**, 113–121 (1995).
10. Rabinowitz, J. D., Beeson, C., Lyons, D. S., Davis, M. M. & McConnell, H. M. *Proc. natn. Acad. Sci. U.S.A.* **93**, 1401–1405 (1996).
11. Karlsson, R., Michaelsson, A. & Mattsson, L. J. *immunol. Meth.* **145**, 229–240 (1991).
12. Corr, M. et al. *Science* **265**, 946–949 (1994).
13. Matsui, K., Boniface, J. J., Steffner, P., Reay, P. A. & Davis, M. M. *Proc. natn. Acad. Sci. U.S.A.* **91**, 12862–12866 (1994).
14. Sykulev, Y. et al. *Immunity* **1**, 15–22 (1994).
15. Sykulev, Y. et al. *Proc. natn. Acad. Sci. U.S.A.* **91**, 11487–11491 (1994).
16. Al-Ramadi, B. K., Jelonick, M. T., Boyd, L. F., Margulies, D. H. & Bothwell, A. L. M. *J. Immun.* **155**, 662–673 (1995).

17. Valitutti, S., Muller, S., Cella, M., Padovan, E. & Lanzavecchia, A. *Nature* **375**, 148–151 (1995).
18. Kupfer, A. & Singer, S. J. *A. Rev. Immun.* **7**, 309–337 (1989).
19. Dustin, M. L., Ferguson, L. M., Chan, P.-Y., Springer, T. A. & Golan, D. E. *J. Cell Biol.* **132**, 465–474 (1996).
20. Kelly, J. M. et al. *Eur. J. Immun.* **23**, 3318–3326 (1993).
21. Arden, B., Clark, S. P., Kabelitz, D. & Mak, T. W. *Immunogenetics* **42**, 501–530 (1995).
22. Koop, B. F. et al. *Genomics* **19**, 478–493 (1994).
23. Bunch, T. A., Grinblat, Y. & Goldstein, L. S. B. *Nucleic Acids Res.* **16**, 1043–1061 (1988).
24. Scorer, C. A., Clare, J. J., McCombie, W. R., Romanos, M. A. & Sreekrishna, K. *Bio/Technology* **12**, 181–184 (1994).
25. Gascoigne, N. R. J. *J. biol. Chem.* **265**, 9296–9301 (1990).
26. Kappler, J., White, J., Kozono, H., Clements, J. & Marrack, P. *Proc. natn. Acad. Sci. U.S.A.* **91**, 8462–8466 (1994).
27. Zhang, W., Young, A. C. M., Imai, M., Nathenson, S. G. & Sacchettini, J. C. *Proc. natn. Acad. Sci. U.S.A.* **89**, 8403–8407 (1992).
28. Ajitkumar, P. et al. *Cell* **54**, 47–56 (1988).
29. Malchiodi, E. L. et al. *J. exp. Med.* **182**, 1833–1845 (1995).

ACKNOWLEDGEMENTS. S.M.A. and P.J.T. contributed equally to this work. We thank F. Carbone for TCR cDNA clones; A. Mellor for the H-2K^b cDNA clone; E. Collins for the β 2m bacterial expression plasmids; W. Studier for the pET bacterial expression plasmids; R. Kubo and J. Moorhead for monoclonal antibodies H57-597 and H28-710; and R. Zamoyka for reading the manuscript. S.M.A. is a fellow of the Concern Foundation for Cancer Research; P.J.T. is a Royal Society University research fellow; and S.C.J. is a special fellow of the Leukemia Society of America. This work was supported by the NIH and MRC, and aided by a NATO collaborative research grant and a Birkbeck College research grant.

CORRESPONDENCE and requests for materials should be addressed to N.R.J.G. (e-mail: gascoigne@scripps.edu).

A human Mad protein acting as a BMP-regulated transcriptional activator

Fang Liu, Akiko Hata, Julie C. Baker*,
Jacqueline Doody, Juan Cárcamo, Richard M. Harland*
& Joan Massagué

Cell Biology Program and Howard Hughes Medical Institute,
Memorial Sloan-Kettering Cancer Center, 1275 York Avenue,
New York, New York 10021, USA

* Department of Molecular and Cell Biology, Division of Biochemistry and
Molecular Biology, University of California, Berkeley, California 94709, USA

THE TGF- β /activin/BMP cytokine family signals through serine/threonine kinase receptors, but how the receptors transduce the signal is unknown. The *Mad* (Mothers against decapentaplegic) gene from *Drosophila*¹ and the related *Sma* genes from *Caenorhabditis elegans*² have been genetically implicated in signalling by members of the bone-morphogenetic-protein (BMP) subfamily. We have cloned *Smad1*, a human homologue of *Mad* and *Sma*. Microinjection of *Smad1* messenger RNA into *Xenopus* embryo animal caps mimics the mesoderm-ventralizing effects of BMP4. *Smad1* moves into the nucleus in response to BMP4. *Smad1* has transcriptional activity when fused to a heterologous DNA-binding domain, and this activity is increased by BMP4 acting through BMP-receptor types I and II. The transactivating activity resides in the conserved carboxy-terminal domain of *Smad1* and is disrupted by a nonsense mutation that corresponds to null mutations found in *Mad* and in the related gene *DPC4*, a candidate tumour-suppressor gene in human pancreatic cancer³. Additionally, we show that *DPC4* contains a transcriptional activation domain. The results suggests that the *Smad* proteins are a new class of transcription factors that mediate responses to the TGF- β family.

To study the involvement of *Mad* family members in BMP signalling, we searched Genbank for human homologues of *Mad* and identified over 60 entries of human expressed-sequence tag (EST) complementary DNA clones. Using three distinct EST clones that have homology to the amino-terminal region of *Mad*, we isolated the corresponding full-length human cDNAs. The closest human homologue of *Mad* that we identified encodes a 465-amino-acid protein designated as *Smad1* (for *Sma* and *Mad*

homologue) (Fig. 1a). All known members of the *Mad* family contain highly conserved N- and C-terminal domains separated by a proline-rich spacer of variable length. A highly conserved segment near the C terminus is the target of mutations in null alleles of *Mad1*, *Sma-2* and *Sma-3* (ref. 2), as well as in *DPC4* alleles from human pancreatic carcinomas with loss of heterozygosity at this locus³. The predicted human *Smad1* protein sequence has this typical domain structure and is 91% identical to *Drosophila* *Mad* over the N- and C-terminal domains (Fig. 1b). The second clone, designated *Smad2*, was later found to be the human homologue of a mouse gene identified by its ability to mimic activin induction of dorsal mesoderm in *Xenopus* (J.C.B. and R.M.H., submitted); the third clone was subsequently identified as *DPC4*.

Given the resemblance of *Smad1* to *Mad* and *Sma-2*, we investigated whether *Smad1* could participate in BMP signalling. To this end, we used a *Xenopus* embryo animal cap assay in which addition of BMP^{4,5} or microinjection of BMP-receptor transcripts^{6–8} potentially ventralizes activin-induced dorsal mesoderm. If *Smad1* participates in BMP signalling, its overexpression would be expected to sensitize cells to endogenous BMP signals and promote the ventralization of activin-induced mesoderm. Indeed, injection of *Smad1* transcripts into animal caps inhibited the induction of the dorsal mesoderm markers *noggin* and *goosecoid* (*gsc*) by activin, did not affect the induction of the general mesoderm marker *Xbrachyury*, and induced the expression of the ventral mesoderm markers *Xwnt8* and *xhox3* (Fig. 2). Injection of a control RNA at the same dosage had no effect on all the markers examined (data not shown). *Smad1* overexpression did not mimic or enhance the growth-inhibitory or transcriptional effects of TGF- β or activin⁹ in R1B/L17 mink lung epithelial cells (data not shown).

To investigate the role of *Smad1* in BMP signalling, we determined its localization in the cell. We introduced a Flag epitope tag to the N terminus of *Smad1*, transfected this construct into COS cells, and immunostained the cells with anti-Flag monoclonal antibody. This revealed that the protein was present in the cytoplasm and nucleus in the absence of BMP and became concentrated in the nucleus in response to BMP4 (Fig. 3). Nuclear localization of *Smad1* was maximal 30–60 min after BMP4 addition. These results indicate that *Smad1* may function downstream of BMP4 and convey BMP4 signals to the nucleus.

In light of these observations, we asked whether *Smad1* has transcriptional activity. To facilitate the analysis of *Smad1* as a potential transcriptional activator, we generated the construct GAL4-*Smad1*(FL) by fusing full-length *Smad1* C-terminally to

the GAL4 DNA-binding domain, GAL4(1-147)¹⁰. As the Smad1 C-terminal domain has some activity in the *Xenopus* animal cap assay (data not shown), we also created a GAL4-Smad1(C) fusion encoding this C-terminal domain. These GAL4 fusions were transfected into R1B/L17 cells together with the reporter con-

struct G1E1BCAT, which contains one GAL4-binding site¹¹. The GAL4-Smad1(FL) construct was inactive in this assay, whereas GAL4-Smad1(C) potently stimulated transcription from the reporter gene (Fig. 4a). Because inactive alleles of *Mad*, *Sma* and *DPC4* result from nonsense mutations in a hot spot near the C terminus¹⁻³, we asked whether a similar mutation could affect the transcriptional activity of Smad1. We generated a GAL4-Smad1 (CA) construct containing the C-terminal domain of Smad1 with a stop codon at the position that normally encodes Gln427. As shown in Fig. 4a, this deletion abolishes the transcriptional activity of Smad1. Figure 4c (left) indicates that the variations in expression of the GAL4 fusion protein are not sufficient to account for their difference in activity. Other Smad proteins also acted as transcriptional activators in these assays. When fused to GAL4(1-147), the C-terminal domain of DPC4 stimulated transcription from a *GAL4* reporter gene, whereas a full-length DPC4 fusion construct was inactive (Fig. 4b), even though both fusion proteins were

a

Hs Smad1	1	-----MNVT	SLFSFTSPAVKRL	LG	19	
Dm Mad	1	MDTDDVESNTSSAMSTL	SLFSFTSPAVKRL	LG	33	
Ce Sma-2	1	-----MINFDGIKKITER	RLK		15	
Hs Smad1	20	WKQGDDEEEKWAEKAVD	ALVKLLKKK	-KGAMEEL	51	
Dm Mad	34	WKQGDDEEEKWAEKAVD	SLVKLLKKR	-KGAIIEEL	65	
Ce Sma-2	16	WKQGDDEEDENWAKKATD	NLMKKLTIKHN	KQALLENL	48	
Hs Smad1	52	EKALSCPGQPSN	-CVTIPRSLDGR	LQVSHRKGL	83	
Dm Mad	66	ERALSCPGQPSK	-CVTIPRSLDGR	LQVSHRKGL	97	
Ce Sma-2	49	EFAALRCQGGQKTE	CVTIPRSLDGR	LQISHRKAL	81	
Hs Smad1	84	PHVIYCRVWRWPD	LQSHHEKLPLE	CC	116	
Dm Mad	98	PHVIYCRVWRWPD	LQSHHEKLPLE	ELCQYIPFSAK	130	
Ce Sma-2	82	PHVIYCRVYRWPD	LQSHHEKLPLE	EDCRFCYESSG	114	
Hs Smad1	117	QKEVCINPYHYKR	VESPL	-VLPPVLVPRHSEYNP	148	
Dm Mad	131	QKEVCINPYHYKR	VESPL	-VLPPVLVPRHSEYAP	162	
Ce Sma-2	115	QKDTICINPYHYKR	VHATG	VLPPVLVPRHSEYKPP	147	
Hs Smad1	149	QHSLLAQFRNLG	QNEPHMPLNAT	FDPSSFQQPNS	181	
Dm Mad	163	GHSLQCFNHVA	---EPSPMHNVS	YSNLSGF	190	
Ce Sma-2	148	CEVPPPTLAKFL	MEMSGSRMPQNVNMANVN	---	177	
Hs Smad1	182	HPFPHSPNS	SYNPSPGSSSTY	PHSPHSTSSDPGS	214	
Dm Mad	191	HSLSSTNTS	VGSFSSVNSNPNS	PDYDSL	217	
Ce Sma-2	178	FTANQFHQYN	PNGIEEMDT	SSQKFDIPPGVPT	208	
Hs Smad1	215	PFQMPAD	TPPPAYLP	PEDEPMTQDGSQPMGDTNMM	247	
Dm Mad	218	-----AGT	TPPPAYS	PEDEGNSNNPNDGMDQLDA	245	
Ce Sma-2	209	-----	-----	-----	208	
Hs Smad1	248	APPLPSEINR	GDVQA	VAYEEPKHMCSTV	YYEELN	280
Dm Mad	246	QM	-----GDVQA	VSYSEPAFWASIA	YYEELN	270
Ce Sma-2	209	-----	CLVPFDK	VWEQEFWATVS	YYEELN	231
Hs Smad1	281	NRVGEAFH	ASSTSV	LVLDGFTDPSNNK	NRFCGL	313
Dm Mad	271	CRVGEVFH	CNNNSVI	LDGFTDPSNNNS	DRCCGLG	303
Ce Sma-2	232	TRYGEQVKV	SSTTIT	DGFTDPCIN	NGSKISLGL	264
Hs Smad1	314	LSNVNRNST	IENRRHIGKGV	HYLVVVG	-GEVY	344
Dm Mad	304	LSNVNRNST	IENRRHIGKGV	HYLVVVT	-GEVY	334
Ce Sma-2	265	FSNVNRNAT	IENRRHIGKGV	LYVRSN	GLSFL	297
Hs Smad1	345	AECLSDS	SAIFVQSR	NCNYHHGFHPT	VCKIP	377
Dm Mad	335	AECLSDS	SAIFVQSR	NCNYHHGFHPT	VCKIP	367
Ce Sma-2	298	AQCFS	SDSAIFVQSR	NCNYINGFHS	TTVVKIANK	330
Hs Smad1	378	CSLKIFNNQ	EFAQLAQSVNH	GFETVYELTKMC		410
Dm Mad	368	CSLKIFNNQ	EFAQLAQSVNH	GFETVYELTKMC		400
Ce Sma-2	331	CSLKIFDMET	FRQLLEDCSR	RGFDAISFDLQKMT		363
Hs Smad1	411	TIRMSFVK	GWGAEYHR	QDVTSTPCWIE	IHLHGP	443
Dm Mad	401	TIRMSFVK	GWGAEYHR	QDVTSTPCWIE	IHLHGP	433
Ce Sma-2	364	FIIRMSFVK	GWGAEYQ	RQDVTSTPCWIE	IHLHAP	396
Hs Smad1	444	LQWLDK	VLTQMGSPHN	PISVS		465
Dm Mad	434	LQWLDK	VLTQMGSPHN	AISSVS		455
Ce Sma-2	397	LAWLD	RVLSTMGPT	PRPISSTIS		418

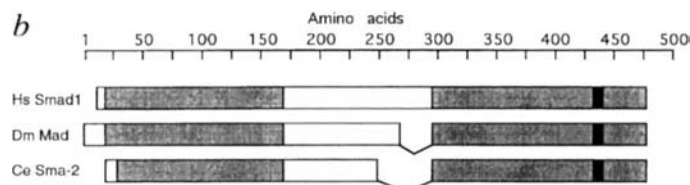


FIG. 1 a, Deduced amino-acid sequence of human Smad1 and alignment with *Drosophila* Mad¹ and *C. elegans* Sma-2 (ref. 2). Identical residues are boxed. Conserved residues mutated in null alleles of *Mad*, *Sma-2*, *Sma-3* or *DPC4* are indicated by dots. b, Diagram of the domain structure of these proteins indicating the conserved N- and C-terminal domains (grey boxes) and the mutational hot spot (black).

METHODS. IMAGE Consortium (LLNL) cDNA clone 163228 was obtained¹⁴ and sequenced. This cDNA clone encompasses codon 24 to ~160 bp within the 3' untranslated region. A riboprobe containing the first 590 bp of clone 163228 was used to screen a λ gt11 human kidney library (Clontech). Of 7.5×10^5 phage plaques screened, 3 positive clones were obtained. In addition to overlapping with clone 163228, these clones provide the missing 23 N-terminal codons and 425 bp of 5' untranslated region (Genbank accession number U59423).

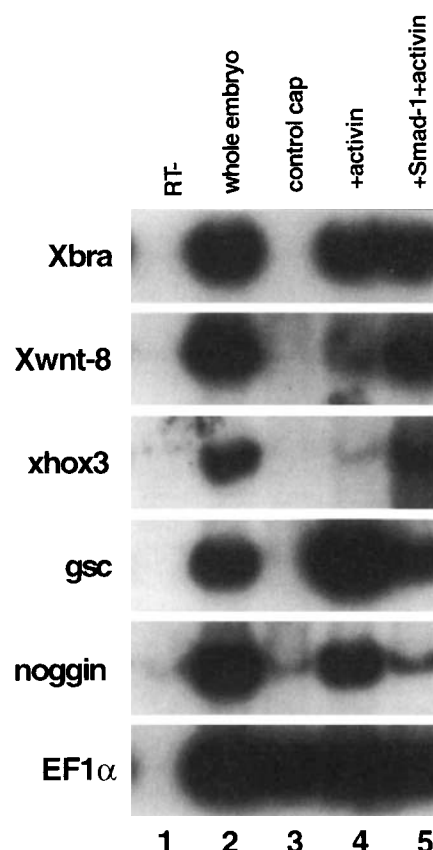


FIG. 2 Smad1 ventralizes mesoderm in *Xenopus* embryos. Animal caps were injected with the Smad1 transcript and analysed for expression of the indicated mesoderm markers at the gastrula stage after treatment with activin at the late blastula stage. *Xbrachyury* (*Xbra*) is a general mesoderm marker at this stage, and *EF1α* was used as a control. Lanes: 1, whole-embryo RNA subjected to polymerase chain reaction with reverse transcriptase (RT-); 2, whole-embryo RNA at a stage equivalent to the animal caps; 3, animal caps incubated in buffer alone; 4, animal caps incubated in activin-conditioned medium; 5, same as lane 4, but the caps were taken from embryos injected with *Smad-1* RNA.

METHODS. Full-length Smad-1 was cloned into the CS2 *Xenopus* expression vector (gift from D. Turner and H. Weintraub) and transcribed *in vitro*. *Smad-1* mRNA (5 ng) was injected into the animal pole of one-cell-stage embryos. Animal caps were cut at stage 9 and cultured to stage 11 in the presence or absence of activin. RNA was collected from ten caps or one embryo using trizol, and subjected to RT-PCR as described^{15,16}. Activin-conditioned medium was derived from COS cells transfected with pKB590 (ref. 17) and used at 1:50 dilution.

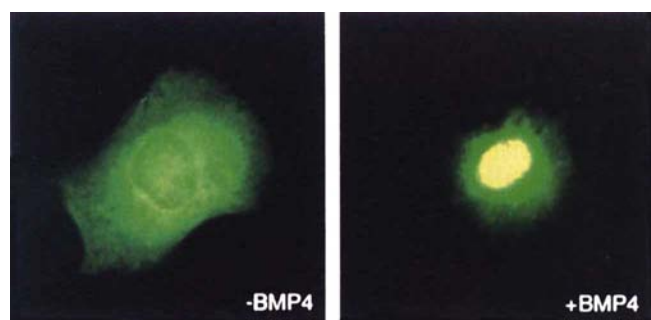


FIG. 3 Smad1 accumulated in the nucleus upon BMP4 stimulation. **METHODS.** pCMV5 vector (6 μ g) encoding Flag-tagged Smad-1 was transfected into COS cells. 48 h later, cells were incubated with 2 nM BMP4 for 30 min. Immunostaining was done with M2 anti-Flag monoclonal antibody (IBI) and FITC-conjugated secondary antibody. Micrographs show representative cells. The predominantly nuclear staining shown in the right hand panel was observed in 22% of the cells with BMP4, but only in 2% of the cells without BMP4.

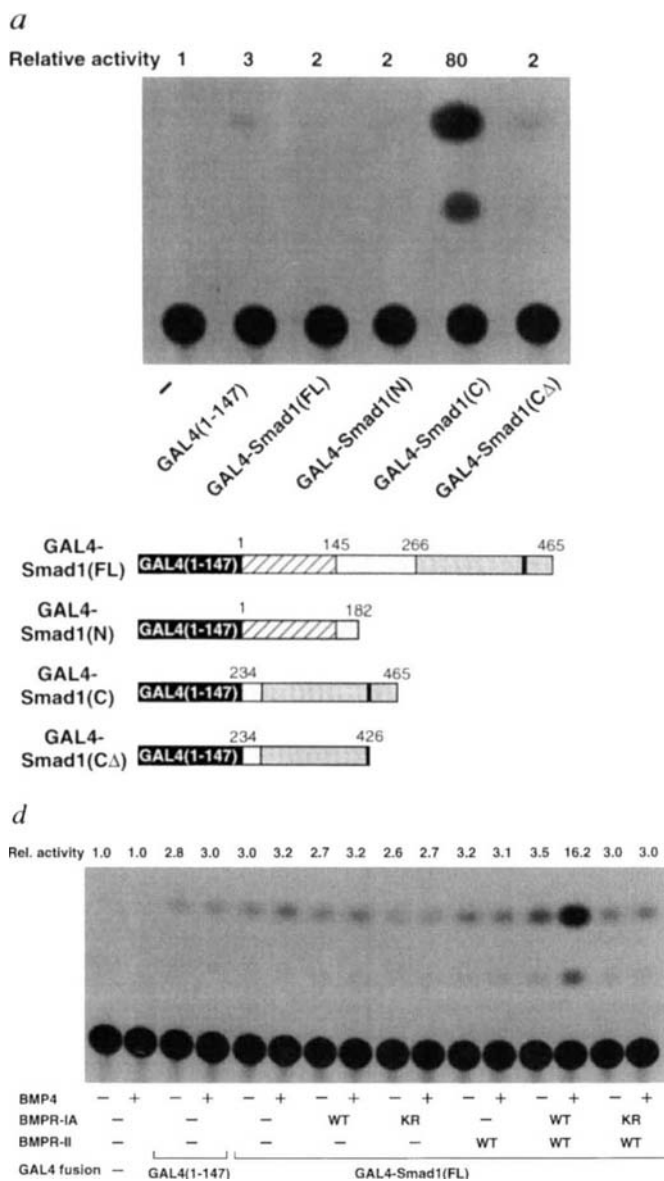
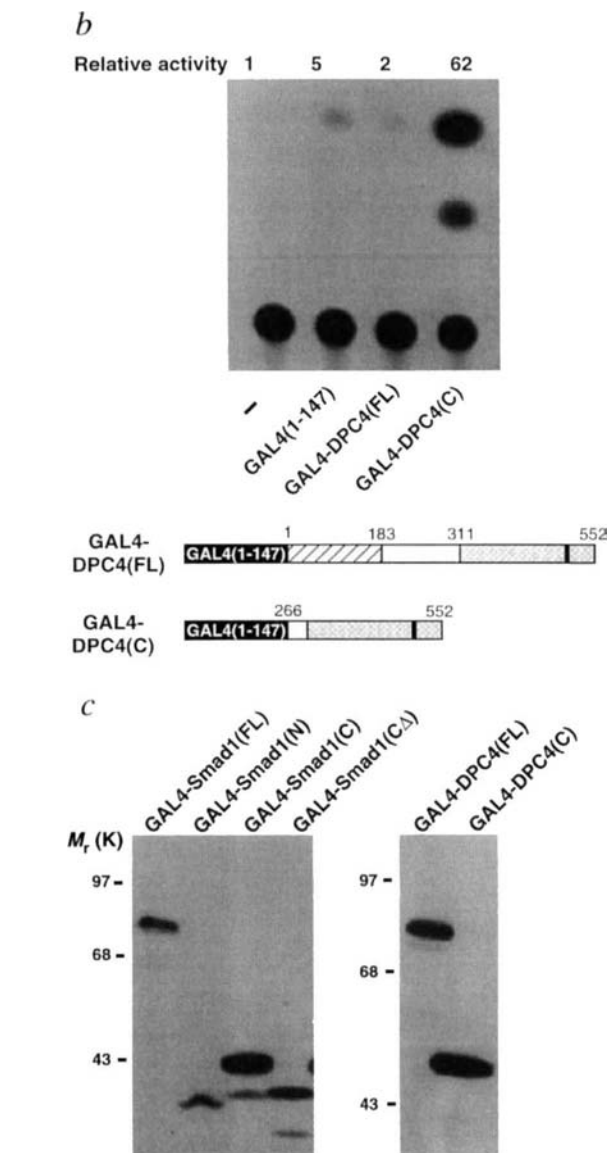


FIG. 4 Transcriptional activity of Smad-1 and DPC-4. **a** and **b**, Structures of the Smad-1 and DPC4-derivatives are shown, together with autoradiograms of chloramphenicol acetyltransferase (CAT) assays. CAT activity was quantified by scintillation counting and is indicated above each lane; CAT activity of the reporter gene alone is arbitrarily defined as unity. **c**, Expression of GAL4 fusion proteins as determined by immunoblot using anti-GAL4 antibody. **d**, Stimulation of the transcriptional activity of full-length Smad-1 by BMP4. GAL4-Smad1(FL) was transfected into R1B/L17 cells together with wild-type (WT) or kinase-defective mutant (KR) BMP receptors. Cells were then treated with BMP4 and assayed for CAT activity. **METHODS.** The appropriate cDNA fragments were cloned in-frame to the GAL4(1-147) cDNA in the pSG424 vector¹⁸. **a**, GAL4-Smad1 (4 μ g) and its derivatives were co-transfected into R1B/L17 cells¹⁹, with 4 μ g of the



GAL4 reporter gene G1E1BCAT¹¹ using the DEAE-dextran technique¹². Forty-eight hours after transfection, cells were collected and CAT activity assayed²⁰. Values represent at least three independent transfection experiments. **b**, DPC4 full-length cDNA was generated by ligating two partial cDNAs available as IMAGE Consortium cDNA clones 15946 and 23275, and a third fragment obtained by PCR from a human liver library (Clontech). 4 μ g GAL4-DPC4(FL) or GAL4-DPC4(C) was transfected into R1B/L17 cells with 4 μ g G5E1BCAT¹¹. Other methods were as for **a**. **c**, COS cells were transfected with 4 μ g GAL4 fusion constructs; 48 h post-transfection, cells were collected, lysed in SDS sample buffer, and identical aliquots were separated on an 8% SDS-polyacrylamide gel and transferred to a nitrocellulose membrane. GAL4 fusion proteins were detected by incubation with an anti-GAL4(1-147) antibody²¹ followed by incubation with [¹²⁵I]protein A. **d**, 2 μ g G1E1BCAT reporter, 0.5 μ g GAL4-Smad1(FL), 3 μ g BMPR-IA (WT or KR) and/or 3 μ g BMPR-II (ref. 12) were cotransfected into R1B/L17 cells as indicated. Twenty-four hours post-transfection, some cultures received 2 nM BMP4; 18 h later, cells were collected and assayed for CAT activity.

comparably expressed (Fig. 4c, right). Similar results were obtained with Smad2 (data not shown). We concluded that the C-terminal domain of Smad proteins has transcriptional activity that can be unmasked by removal of the N-terminal domain.

The inability of the full-length Smad1 to stimulate transcription suggested that its activation may be regulated. Reasoning that activation of the full-length Smad1 may require a BMP signal, we cotransfected this construct with BMP receptors types I and II in R-1B/L17 cells and analysed the effect of the ligand. Responsiveness of R-1B/L17 cells to BMP requires transfection of these two receptor serine/threonine kinases^{12,13}. As shown in Fig. 4d, BMP4 addition to cells transfected with either the BMP receptor I (BMPI-IA) or receptor II (BMPI-II) did not activate GAL4-Smad1(FL). In contrast, BMP4 addition to cells cotransfected with both BMPI-IA and BMPI-II increased the transcriptional activity of GAL4-Smad1(FL) ~5-fold relative to the same cells without BMP4 (Fig. 4d). Similar results were obtained with BMP2 or the receptor isoform BMPI-IB (data not shown). This effect is dependent on the kinase activity of the receptors because kinase-defective mutant versions of BMPI-IA (Fig. 4d) or BMPI-II (data not shown) failed to mediate this BMP4 response. Controls showed that BMP4 does not increase the expression of the GAL4-Smad1(FL) protein, and cotransfection of BMPI-IA and BMPI-II, with or without BMP4 addition, does not increase the activity of GAL4(1-147) (data not shown).

In conclusion, we have shown that Smad1 is concentrated in the nucleus upon BMP stimulation and that it can mediate BMP signals. Moreover, we have shown that Smad1 and other Smad proteins, including DPC4, encode transcriptional activators. This activity is located in their C-terminal domain and is unmasked upon removal of the N-terminal domain. We further demonstrate that the transcriptional activity of Smad1 can be stimulated by BMP-receptor-mediated signals. How these signals regulate the transcriptional activity of Smad1 is not clear. We have observed that BMP stimulation increases Smad1 phosphorylation (data not shown), but it remains to be determined whether phosphorylation regulates the transcriptional activity of Smad1. The finding that Smad1 and other Smad proteins are transcriptional activators is a step towards a better understanding of gene regulation by the TGF- β family.

Note added in proof: *Smad1* is identical to the recently published *Madrl* (ref. 22) and is the homologue of *Xenopus Xmad1* (ref. 23). □

Received 9 May; accepted 23 May 1996.

1. Sekelsky, J. J., Newfeld, S. J., Raftery, L. A., Chartoff, E. H. & Gelbart, W. M. *Genetics* **139**, 1347-1358 (1995).
2. Savage, C. et al. *Proc. natn. Acad. Sci. U.S.A.* **93**, 790-794 (1996).
3. Hahn, S. A. et al. *Science* **271**, 350-353 (1996).
4. Dale, L., Howes, G., Price, B. M. J. & Smith, J. C. *Development* **115**, 573-585 (1992).
5. Jones, C. M., Lyons, K. M., Lapan, P. M., Wright, C. V. E. & Hogan, B. L. M. *Development* **115**, 639-647 (1992).
6. Graff, J. M., Thies, R. S., Song, J. J., Celeste, A. J. & Melton, D. A. *Cell* **79**, 169-179 (1994).
7. Suzuki, A. et al. *Proc. natn. Acad. Sci. U.S.A.* **91**, 10255-10259 (1994).
8. Harland, R. M. *Proc. natn. Acad. Sci. U.S.A.* **91**, 10243-10246 (1994).
9. Cárcamo, J. et al. *Molec. cell. Biol.* **14**, 3810-3821 (1994).
10. Ptashne, M. *Nature* **335**, 683-689 (1988).
11. Lillie, J. W. & Green, M. R. *Nature* **338**, 39-44 (1989).
12. Liu, F., Ventura, F., Doody, J. & Massagué, J. *Molec. cell. Biol.* **15**, 3479-3486 (1995).
13. Rosenzweig, B. L. et al. *Proc. natn. Acad. Sci. U.S.A.* **92**, 7632-7636 (1995).
14. Lennon, G. G., Auffray, C., Polymeropoulos, M. & Soares, M. B. *Genomics* (in the press).
15. Rupp, R. A. & Weintraub, H. *Cell* **65**, 927-937 (1991).
16. Wilson, P. A. & Melton, D. A. *Curr. Biol.* **4**, 676-686 (1994).
17. Lamb, T. M. et al. *Science* **262**, 713-718 (1993).
18. Sadowski, I. & Ptashne, M. *Nucleic Acids Res.* **17**, 7539 (1989).
19. Wrana, J. L., Attisano, L., Wieser, R., Ventura, F. & Massagué, J. *Nature* **370**, 341-347 (1994).
20. Gorman, C. M., Moffat, L. F. & Howard, B. H. *Molec. cell. Biol.* **2**, 1044-1051 (1982).
21. Liu, F. & Green, M. R. *Cell* **61**, 1217-1224 (1990).
22. Hoodless, P. A. et al. *Cell* **85**, 489-500 (1996).
23. Graff, J. M., Bansal, A. & Melton, D. A. *Cell* **85**, 479-487 (1996).

ACKNOWLEDGEMENTS. We thank M. Bosenberg for insightful suggestions, G. Lennon and F. Bouillaud for EST cDNA clones, the Genetics Institute for BMP4, and I. Reynisdottir for assistance. This work was supported by grants from the NIH to R.M.H. and to J.M. and to the Memorial Sloan-Kettering Cancer Center. F.L. and J.C.B. are fellows of the Jane Coffin Childs Memorial Fund for Medical Research. A.H. is a research associate and J.M. is an investigator of the Howard Hughes Medical Institute.

CORRESPONDENCE and requests for materials should be addressed to J.M. (e-mail: j-massague@ski.mskcc.org).

Intersubunit rotation in active F-ATPase

D. Sabbert, S. Engelbrecht & W. Junge

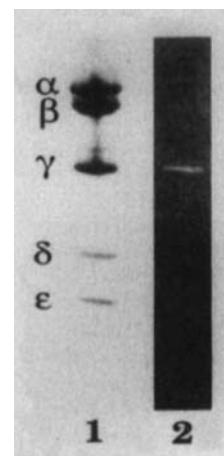
Abteilung Biophysik, Fachbereich Biologie/Chemie, Universität Osnabrück, D-49069 Osnabrück, Germany

THE enzyme ATP synthase, or F-ATPase, is present in the membranes of bacteria, chloroplasts and mitochondria. Its structure is bipartite, with a proton-conducting, integral membrane portion, F_0 , and a peripheral portion, F_1 . Solubilized F_1 is composed of five different subunits, $(\alpha\beta)_3\gamma\delta\epsilon$, and is active as an ATPase^{1,2}. The function of F-ATPase is to couple proton translocation through F_0 with ATP synthesis in F_1 (ref. 3). Several lines of evidence support the spontaneous formation of ATP on F_1 (refs 4, 5) and its endergonic release⁶ at cooperative and rotating (or at least alternating) sites⁷. The release of ATP at the expense of protonmotive force might involve mechanical energy transduction from F_0 into F_1 by rotation of the smaller subunits (mainly γ) within $(\alpha\beta)_3$, the catalytic hexagon of F_1 as suggested by electron microscopy⁸, by X-ray crystal structure analysis⁹ and by the use of cleavable crosslinkers¹⁰. Here we record an intersubunit rotation in real time in the functional enzyme by applying polarized absorption relaxation after photobleaching to immobilized F_1 with eosin-labelled γ . We observe the rotation of γ relative to immobilized $(\alpha\beta)_3$ in a timespan of 100 ms, compatible with the rate of ATP hydrolysis by immobilized F_1 . Its angular range, which is of at least 200 degrees, favours a triple-site mechanism of catalysis^{7,11}, with γ acting as a crankshaft in $(\alpha\beta)_3$. The rotation of γ is blocked when ATP is substituted with its non-hydrolysable analogue AMP-PNP.

We investigated intersubunit rotation in the active F_1 -ATPase CF_1 from spinach chloroplast. The enzyme was purified and modified covalently by eosin-5-maleimide on subunit- γ . Figure 1 shows the specificity of labelling, as indicated by SDS-PAGE. The labelled enzyme was immobilized on Sephadex DEAE-A50 (ref. 12). We used polarized absorption relaxation after photobleaching

FIG. 1 Covalent modification of chloroplast F_1 -ATPase, CF_1 , with eosin-5-maleimide. SDS-PAGE of CF_1 (lane 1), eosin-5-maleimide labelled CF_1 (lane 2). Samples were electrophoresed in Pharmacia Phast gradient gels (8-25%), and silver-stained (lane 1)¹³ or photographed under UV light (lane 2); 0.3 μ g protein was loaded per lane.

METHODS. CF_1 was prepared by anion-exchange chromatography of EDTA extracts from spinach chloroplasts²⁰. The specific activity of the enzyme was 25 U mg⁻¹ (measured in 50 mM Tris-HCl, 5 mM ATP, 2 mM MgCl₂, 30 mM 1-O-n-octyl- β -D-glucopyranoside (Sigma), 10 mM Na₂SO₃, with 5 min incubation at 37 °C); the reaction was terminated with 0.5 M trichloroacetic acid. Phosphate concentration was determined according to ref. 21. Specific labelling of CF_1 (5 μ M) was achieved at room temperature for 10 min in the dark at pH 7.0 (50 mM MOPS-NaOH) with 50 μ M eosin-5-maleimide (Molecular Probes); the reaction was terminated with 1 mM N-acetylcysteine. Labelling decreased the specific Mg-ATPase activity from 25 to 20 U mg⁻¹. The reaction with eosin-5-maleimide occurred in the γ -subunit. Under these conditions, eosin-5-maleimide is expected to react only with the so-called 'dark site' in spinach chloroplast γ , namely residue $CF_1\gamma$ (C322)^{16,17}. The stoichiometric proportion of bound dye (extinction coefficient at 530 nm, 10⁵ M⁻¹ cm⁻¹) over enzyme was >50% (mol per mol). Labelled CF_1 was immobilized on Sephadex DEAE-A50. Unbound CF_1 was removed by washing. Upon immobilization the specific activity of eosin-labelled CF_1 was decreased ~10-fold to 2 $\pm$ 1 U mg⁻¹ (ATP-turnover number, 14 $\pm$ 7 s⁻¹).



(PARAP) to detect slow rotational motion of eosin-labelled subunit- γ relative to immobilized $(\alpha\beta)_3$. The results of PARAP are illustrated in Fig. 2. The decay of the polarization anisotropy parameter, r , reflects only and directly the rotational motion of the chromophore (Fig. 2, bottom)^{13,14}. If the same non-degenerate optical transition is both excited and interrogated, as in our experiments with eosin, the theoretically expected initial value of r is 0.4 (ref. 14). If the chromophore rotates fully around all three cartesian axes, r decays to zero. If the rotation is more restricted (dimensionally or by boundaries), r may decay to a value greater than zero. If the linear oscillator rotates uniaxially in the plane of a circle, r decays to a value of 0.1 (ref. 15). Broadly, the rate constants of the decay of r are indicative of the 'velocity' of rotational motion, and its steady value is indicative of restrictions to full rotational freedom.

Figure 3 shows the value of the anisotropy parameter as a function of the time after firing the laser flash. CF₁ was labelled with eosin-5-maleimide at the C terminus of subunit- γ at Cys322 (refs 16,17). The initial value of the anisotropy was $r(t=0) = 0.1$, smaller than the theoretical maximum. This was due to a very fast spinning of the chromophore around its single bonds (nanosecond timescale). r decayed to a stationary value of $r(t \rightarrow \infty) = 0.018$ (Fig. 3, middle trace). This decay was only observed under conditions of hydrolysis: that is, in the presence of ATP and MgCl₂ to drive the enzyme, and of *n*-octyl- β -D-glucopyranoside at 37 °C to activate it. Its time constant was of the order of 100 ms (a theoretical analysis, see fit line in Fig. 3, will be presented elsewhere). If, on the other hand, ATP was substituted by its non-hydrolysable analogue AMP-PNP, the slow decay of r was blocked (upper trace). When CF₁ was free in aqueous buffer solution, r was

zero (Fig. 3, lower trace). The reduction from 0.4 to zero was due to the very rapid rotational diffusion of CF₁ in solution (rotational correlation time ~ 200 ns; ref. 18). Neither the librational motion of the dye, nor the rapid rotational diffusion of the protein were time-resolved in these experiments. The possibility that the activity-linked decay of the polarization anisotropy could be due to a non-immobilized subset of CF₁-molecules is eliminated by the constancy of r in the presence of AMP-PNP.

Figure 4a shows a space-filling model of $(\alpha\beta)_3$, with the front region removed to expose subunit- γ , displayed as a ribbon. The α -helix extending almost to the top terminates at the C terminus of γ . In the chloroplast enzyme, its penultimate residue is Cys 322, which reacts with eosin-5-maleimide, here illustrated by its transition moment as rotating on a circle (Fig. 4b). As already emphasized⁹, the crystal structure of mitochondrial F₁ATPase leaves little freedom for the librational motion of the long axis of γ . Although in a dynamic situation, the bearing for γ in $(\alpha\beta)_3$ may be looser than suggested by the stationary crystal picture, we assume that the angle ρ in Fig. 4b is small and negligible (a quantitative analysis will be presented elsewhere). Two rotational degrees of freedom remain: first, rapid brownian rotation around the single linker bonds(s) of the chromophore with the rotation axis inclined at an angle θ against the long axis of γ , and second, the rotation of γ around its long axis over an angular interval ϕ (Fig. 4b).

We interpreted the steady level of r observed under ATP hydrolysis, namely $r(t \rightarrow \infty) = 0.018$, in terms of Wahl's¹⁵ theory of angularly restricted, uniaxial brownian rotation. Adapting his equation (50) to our case, we obtained an expression linking the observed value of $r(t \rightarrow \infty)$ with the two angles θ and ϕ (Fig. 4c legend). A contour plot of $r(t \rightarrow \infty, \theta, \phi)$ over the

FIG. 2 Polarized absorption relaxation after photobleaching (PARAP). Two beams of linearly polarized light, the laser flash and the measuring beam, impinge at right angles on the cuvette containing an originally isotropic sample of labelled CF₁. The exciting laser flash causes absorption transients which are recorded by the continuous measurement light.

METHODS. The sample was excited by a linearly polarized laser flash (Q-switched Nd-YAG at 532 nm for 5 ns) at non-saturating energy (0.4 mJ mm⁻²). The flash excites an oriented subset of chromophore molecules with more or less parallel orientation of their transition moments to the $\mathbf{E}$ vector of the laser light. With about 70% probability, eosin crosses over from its excited singlet state into the triplet state²² and decays into the electronic ground state in <1 ms. Only a minority of molecules (about 0.2% in the absence of oxygen) is irreversibly bleached. The bleaching of this minority is used to probe the putative slow rotation of γ . We measured absorption transients at 520 nm, attributable to the same $\pi \rightarrow \pi^*$ transition that is excited by the laser flash at 532 nm. Because of the orientational photoselection by the linearly polarized flash, the extent of absorption transients, which are detected by a linearly polarized, continuous measuring light, differs according to whether the $\mathbf{E}$ vector of the interrogating beam is polarized in parallel ($\Delta A_{\parallel}$) or perpendicularly ($\Delta A_{\perp}$) to the one of the exciting flash. If the chromophores are fixed, the difference is constant. It decays if they rotate, as illustrated in the lower portion of the figure. Usually rotational motion is analysed in terms of the decay of the anisotropy parameter, r (ref. 14): $r(t) = \Delta A_{\parallel} - \Delta A_{\perp} / \Delta A_{\parallel} + 2\Delta A_{\perp}$. If the same non-degenerate optical transition is both excited and interrogated, as in our experiments with eosin, the theoretically expected initial figure of r is 0.4 (ref. 14). With immobilized eosin in polyacrylamide we observed a value of 0.37 in the set-up shown in Fig. 2, conforming with this expectation.

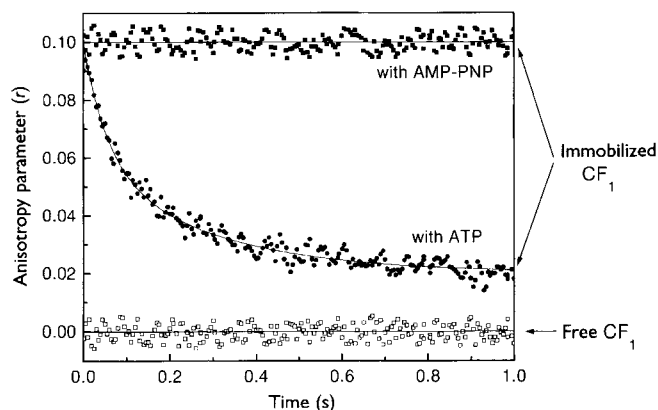
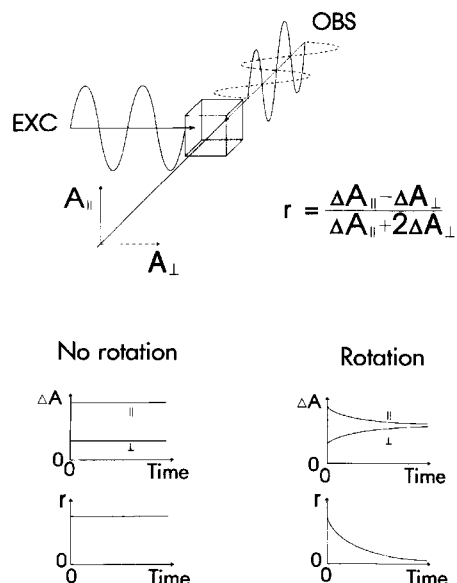


FIG. 3 The anisotropy parameter r as function of the time after a laser flash at time zero to an isotropic ensemble of eosin-labelled CF₁. The sample, 0.5 mg CF₁ immobilized per ml Sephadex DEAE-A50 (suspended in Tris-HCl, pH 7.8) in a quartz cuvette (5 × 5 × 5 mm³), was treated with 2 mM MgCl₂, 30 mM *n*-octyl- β -D-glucopyranoside and 5 mM ATP. Addition of β -D-glucose and glucose oxidase/catalase (10/10 μ g; Sigma), rendered the sample anaerobic, as verified by the phosphorescence lifetime of eosin. After 15 min incubation, the cuvette was heated to 37 °C to stimulate ATP hydrolysis. The sample was excited by repetitive laser flashes at non-saturating energy (density, 0.4 mJ mm⁻²). Polarized absorption transients were recorded and averaged digitally. A new sample was introduced after 20% irreversible bleaching. Data points are the average of 2,000 observations. CF₁ was labelled on subunit- γ by eosin-5-maleimide (Fig. 2, lane 3). Top trace: immobilized CF₁ with 5 mM 5'-adenylylimidodiphosphate (AMP-PNP); middle immobilized CF₁ with ATP; bottom: CF₁ free in solution with ATP added.

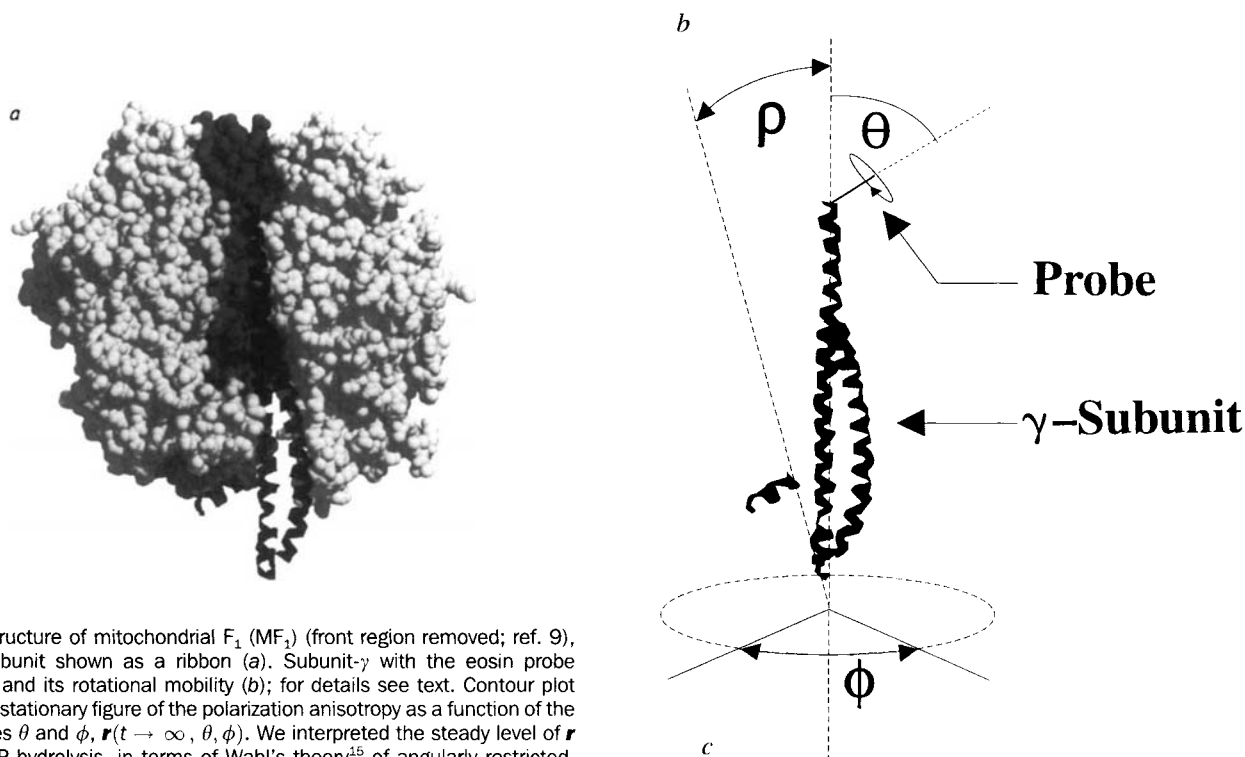


FIG. 4 Structure of mitochondrial F₁ (MF₁) (front region removed; ref. 9), with γ-subunit shown as a ribbon (a). Subunit-γ with the eosin probe attached and its rotational mobility (b); for details see text. Contour plot (c) of the stationary figure of the polarization anisotropy as a function of the two angles θ and ϕ, $r(t \rightarrow \infty, \theta, \phi)$. We interpreted the steady level of r under ATP hydrolysis, in terms of Wahl's theory¹⁵ of angularly restricted, uniaxial brownian rotation. Adapting his equation (50) to our case, we obtain the following expression for the extent of the steady level of r :

$$r(t \rightarrow \infty, \theta, \phi) = r(t = 0) \cdot \left[K_1 \frac{\sin^2(\phi/2)}{(\phi/2)^2} + K_2 \frac{\sin^2 \phi}{\phi^2} + K_3 \right]$$

with $K_1 = 3 \sin^2 \theta \cos^2 \theta$, $K_2 = 0.75 \sin^4 \theta$, $K_3 = 0.25(3 \cos^2 \theta - 1)^2$, $r(t = 0) = 0.1$. The angular domain of the uniaxial rotation, ϕ, and the inclination of the binding axis of the label to the long axis of γ, θ, both of which enter into these equations, are shown in b. The thick line in c denotes all sets of ϕ and θ that are compatible with the experimental value of $r = 0.018$ (hydrolysing enzyme), as obtained for the rotation of γ relative to immobilized $(\alpha\beta)_3$.

plane spanned out by θ and ϕ is shown in Fig. 4c. It shows that the observed steady value of 0.018 implied at least 200 degrees (for θ = 60°) of rotational freedom of γ around its long axis (angle ϕ). This was a lower limit for two reasons. First, if θ was different from 60°, the rotational freedom (angle ϕ) of γ was larger; and second, a subset of enzymes where γ was also immobilized would have increased the apparent value of $r(t \rightarrow \infty)$ over the true value of the complement set of active enzyme molecules.

In conclusion, the hydrolysis of ATP by the chloroplast ATPase CF₁ is accompanied by the rotational motion of subunit-γ around its long axis within and relative to the hexagonal array of $(\alpha\beta)_3$, which contains the three catalytic nucleotide-binding sites. To our knowledge this is the first time that intersubunit rotation over a large angular domain has been resolved in real-time. Its rate conformed with the rate of the catalytic turnover, and its angular range extended over at least 200 degrees. These observations corroborate the functional relevance of intersubunit rotation in ATP hydrolysis. How proton translocation through the membrane portion of ATP synthase, F_o, generates the torque to rotate γ within $(\alpha\beta)_3$ of F₁ remains to be established. □

- Boyer, P. D., Cross, R. L. & Momsen, W. *Proc. natn. Acad. Sci. U.S.A.* **70**, 2837–2839 (1973).
- Grubmeyer, C., Cross, R. L. & Penefsky, H. S. *J. biol. Chem.* **25**, 12092–120100 (1982).
- Penefsky, H. S. *Proc. natn. Acad. Sci. U.S.A.* **82**, 1589–1593 (1985).
- Boyer, P. D. *Biochim. biophys. Acta* **1140**, 215–250 (1993).
- Gogol, E. P., Johnston, E., Aggeler, R. & Capaldi, R. A. *Proc. natn. Acad. Sci. U.S.A.* **87**, 9585–9589 (1990).
- Abrahams, J. P., Leslie, A. G. W., Lutter, R. & Walker, J. E. *Nature* **370**, 621–628 (1994).
- Duncan, T. M., Bulygin, V. V., Zhou, Y., Hutcheon, M. L. & Cross, R. L. *Proc. natn. Acad. Sci. U.S.A.* **92**, 10964–10968 (1995).
- Weber, J., Wilke-Mounts, S. & Senior, A. E. *J. biol. Chem.* **269**, 20462–20467 (1994).
- Junge, W., Schaffernicht, H. & Nelson, N. *Biochim. biophys. Acta* **462**, 73–85 (1977).
- Perrin, F. *Le Journal de Physique et le Radium* **7**, 1–11 (1936).
- Jablonski, A. Z. *Physik* **96**, 236–246 (1935).
- Wahl, P. *Chem. Phys.* **7**, 210–219 (1975).
- Moroney, J. V., Fullmer, C. S. & McCarty, R. E. *J. biol. Chem.* **259**, 7281–7285 (1984).
- Miki, J., Maeda, M., Mukohata, Y. & Futai, M. *FEBS Lett.* **232**, 221–226 (1988).
- Wagner, R. & Junge, W. *Biochemistry* **21**, 1890–1899 (1982).
- Engelbrecht, S., Schürmann, K. & Junge, W. *Eur. J. Biochem.* **179**, 117–122 (1989).
- Engelbrecht, S., Lill, H. & Junge, W. *Eur. J. Biochem.* **160**, 635–643 (1986).
- LeBel, D., Poirier, G. G. & Beaudoin, A. R. *Analyt. Biochem.* **85**, 86–89 (1978).
- Parker, C. A. & Hatchard, C. G. *Trans. Faraday Soc.* **57**, 1894–1909 (1961).

ACKNOWLEDGEMENTS. We thank R. Wagner for advice, H. Kenneweg and M. Roelevink for technical assistance, H. Lill, A. Mulikidjanian and J. Walker for discussion. This work was supported by the Deutsche Forschungsgemeinschaft, the Land Niedersachsen and the Fonds der Chemischen Industrie.

CORRESPONDENCE and requests for materials should be addressed to W.J. (e-mail: junge@uni-osnabrueck.de).

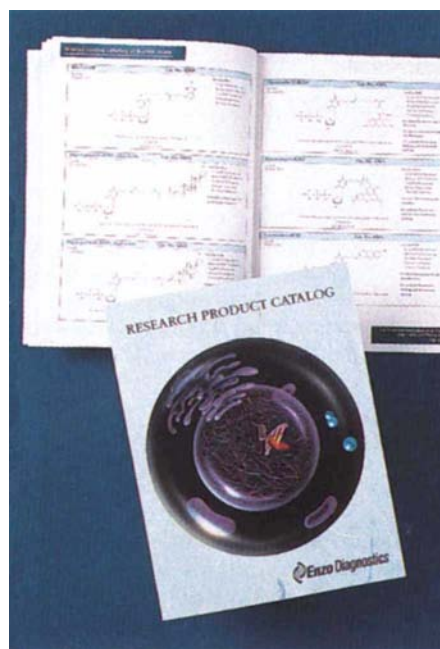
Received 12 March; accepted 24 April 1996.

- Senior, A. E. *A. Rev. Biophys. biophys. Chem.* **19**, 7–41 (1990).
- Walker, J. E. *The Biochemist* **16**, 31–351 (1994).
- Mitchell, P. *Nature* **191**, 144–148 (1961).

The details on DNA

Nature's Internet site also offers product review at www.nature.com and www.america.nature.com — users can respond to product information through the electronic reader response, not only the current issue, but also to past issues.

ENZO now offers its research catalogue for **nucleic acid labelling**, which includes a complete line of biotin-, digoxigenin- and fluorescent-modified nucleotides. These nucleotides are offered in easy-to-use, modular systems that allow the researcher to produce nucleic acid probes by such procedures as nick translation, terminal labelling, random prim-



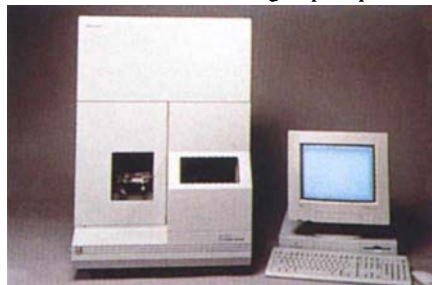
Enzo catalogue for nucleic acid labelling.

ing and RNA labelling, as well as the company's OligoBridge labelling system. Detection reagents include the Detek signal generating systems for rapid colorimetric detection in both *in situ* and membrane hybridization procedures. Enzo's MaxSense membrane hybridization and detection systems offer the scientist all the reagents required for nonradioactive Southern, northern or dot-blot analysis. For *in situ* hybridization, the company has developed complete modular systems for both *in situ* hybridization and detection. PathoGene kits are available for the *in situ* detection of important infectious agents, including HPV, CMV, HBV, EBV, adenovirus, *Chlamydia* and HSV. Kits for the detection of HIV, *Mycobacterium tuberculosis* complex and both hepatitis B surface antigen and core antigen are also available. These labelling and detection sys-

tems have been designed for both the experienced molecular biologist and the first-time user.

Reader Enquiry No. 100

Perkin-Elmer has introduced the ABI Prism 310 genetic analyser to bring **automated sequencing, sizing and quantification of nucleic acids** to more life science laboratories with greater ease of use and lower cost. This system utilizes the company's multicolour fluorescent labels, capillary electrophoresis (CE) and software for the collection and analysis of fluorescent signals. All functions, from sample and gel loading to data collection and analysis, are automated. The analyser should be well suited for small laboratories that previously could not afford sophisticated fluorescent technology. It is said to eliminate time-consuming tasks, such as pouring acrylamide and agarose slab-gels and loading samples onto the gel for electrophoresis — users can begin receiving data in less than half an hour, according to the manufacturer. After a capillary has been installed, the gel pump loads



Perkin-Elmer's ABI genetic analyser.

the capillary with either DNA sequencing or GeneScan polymer. Preformulated polymers are designed to eliminate much of the guesswork involved with preparing gel materials. As the analyser detects multiple fluorescent labels, a discretely labelled size standard can be run along with the PCR samples. This allows fragment mobility rates to be normalized by the GeneScan software, ensuring accurate, precise results. With multiple-dye detection, the analyser collects signals for all sequencing reactions simultaneously, removing electrophoretic variables and allowing the DNA sequencing software to call base sequences reliably.

Reader Enquiry No. 101

Cloning and transfection

National Biosciences has announced the release of the novel 3G Blue **cloning kit**. This prokaryotic cloning system is designed to assist researchers in cloning their PCR products quickly, efficiently and economically. The kit is based on the creation of heteroduplex PCR products with a 5'-GGG overhang. The heteroduplex is easily generated with sets of related PCR primer pairs supplied by the company. The transforming efficiency of the kit is said to be an improvement of five to ten times. The 5'-GGG/3'-CCC duplex formed between the PCR product and the vector is more stable than the single base pairing formed in vectors utilizing the T/A principle. This improved cloning efficiency should prove to be valuable when cloning large PCR products, for example inserts of 5 kb and above, or when the insert is in small quantity. Moreover, this kit does not rely on the terminal 'extendase' activity of the *Taq* enzyme, high fidelity enzymes, such as *Pfu* polymerase, can be used. For those who prefer to utilize *Taq*, the kit's high efficiencies indicate that stringent PCR conditions can be used to produce a more faithful reproduction of the template. The company's subcloning service will subclone PCR products into any vector of interest for researchers lacking the time to subclone their PCR products and, if preferred, NBI will sequence the insert.

Reader Enquiry No. 102

Dosper, a high-performance **DNA transfection reagent** from Boehringer Mannheim, is designed to perform transfections of eukaryotic cells in both serum-free and serum-supplemented culture media. It is also useful in the transfection of molecules, such as RNA and oligonucleotides. According to the manufacturer, this reagent yields improved transfection efficiency through its polycationic liposome-mediated transfection technology. Cationic liposomes interact spontaneously with DNA resulting in the formation of liposome/polynucleotide complexes. Furthermore, the polycationic nature increases DNA binding efficiency, allowing it to carry more DNA into eukaryotic cells. The reagent transfects both common and more complex cell lines and, being less cytotoxic than other transfection reagents, allows

more transfected cells to survive. Dosper is supplied as a convenient, ready-to-use solution and is packaged under argon to prevent auto-oxidation.

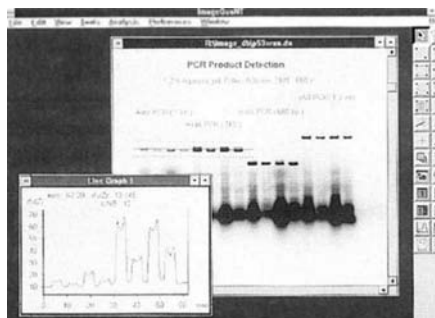
Reader Enquiry No. 103

Oligo operations

For unattended, **postsynthesis oligonucleotide processing**, the Savant OligoPrep automates cleavage, deprotection and concentration steps in synthesis protocols. During the incubation stage, this compact, integrated, benchtop system cleaves and deprotects oligonucleotide samples, then, during concentration, samples are dried into a compact pellet. Both the incubation time (from 1 to 20 hours) and the concentration time (from 1 to 9.75 hours) are fully programmable. The system features a built-in, oil- and maintenance-free vacuum pump, a Teflon-coated concentration chamber and a corrosion-resistant vapour path. The ammonia-neutralizing post trap eliminates NH_3 vapours in the laboratory and provides environmentally safe solvent disposal. There is a choice of two systems: the OP120 for processing up to 12 samples and the high-capacity OP420 for processing up to 42 samples.

Reader Enquiry No. 104

Amersham has announced the availability of the **Vistra fluorescence 5' oligolabelling kit**. This should prove suitable for PCR primer labelling and gel shift assays.



Vistra's fluorescent 5' oligopriming.

The kit's novel chemistry attaches a single fluorescein dye molecule to the 5' end of any oligo quickly and easily. Designed for use with the FluorImager SI, oligonucleotide-based assays require no film exposures or darkroom processing and provide rapid, quantitative results. The company also offers a datasheet describing this new kit.

Reader Enquiry No. 105

Genetic Medizin Laboratories has recently developed a new line of **antisense oligos** with an improvement in their resistance to nucleolytic activities, as well as improved cellular uptake. According to the manufacturer, phosphorothioate antisense oligomers containing four consecutive Gs gave a non-sequence-specific growth inhibition of certain cells. In order to bring about improved antisense activity, the company has introduced a programme requiring synthesis of different kinds of derivatives: second-generation oligonucleotide analogues with novel modifications at both the 3' and 5' ends. Additional positions along the phosphate backbone, which do not contain sulphur, retain full activity. There are also conjugate groups attached to improve cellular uptake and conjugates that have RNase-like activity. The company intends to combine these structural aspects into antisense custom-designed oligos available for its customers.

Reader Enquiry No. 106

Hybridization and detection

Signet's new Rembrandt series of **in situ hybridization and detection systems** are designed to detect viral DNA on formalin-fixed, paraffin-embedded tissue sections, frozen sections and cytology specimens by using the highly sensitive and specific *in situ* hybridization (ISH) technique. According to the manufacturer, ISH is preferable to other laboratory viral detection methods because it enables the early detection of specific DNA sequences without the loss of essential morphological details. The series comprises kits for panHPV (screening), HPV (typing), CMV, HSV 1/2, and EBV, either horseradish peroxidase/AEC or alkaline phosphatase/NBT are available. The system offers a choice of both biotin-

and digoxigenin-labelled probes. DNA probes are labelled using a universal linkage system; this non-enzymatic labelling method is said to ensure enhanced performance and reproducibility. Kits contain ready-to-use reagents in dropper bottles and are supplied complete with all components needed for the ISH procedure: hybridization reagents, probe(s), slides and coverslips.

Reader Enquiry No. 107

Amersham has launched **DYEnamic fluorescence energy transfer primers**, which use two fluorescent labels to generate up to a stated 12-fold increase in fluorescence intensity. According to the manufacturer, this should enable reliable (95 per cent accuracy) read-lengths of up to 1,000 bases (98.5 per cent accuracy for up to 550 bases) to be obtained in conjunction with Thermo Sequenase in high-throughput sequencing labs. These primers can produce up to 75-100 more bases of high-quality data per sequence. The increased signal intensity has the added benefit of reducing the amount of template needed to generate sequence data. DYEnamic primers have been developed specifically to exploit the capabilities of the current generation of automated fluorescence sequencers. These instruments use an argon laser to excite the fluorescent label on a DNA primer and then measure emitted light. By using four dyes with different emission characteristics, the four bases of DNA can be represented by four different colours, facilitating better base identification. The primers use resonance fluorescence energy transfer from a donor dye (called FAM) to an acceptor dye (one of four different dyes: FAM, REG, TAMRA and ROX). The absorption spectrum of the acceptor dye overlaps the emission spectrum of the donor dye; this effectively amplifies the overall signal with the resultant increase in base-calling accuracy. These primers are available as -40 M13 forward or -28 M13 reverse four-dye primer sets, each primer set having matched electrophoretic mobility. Three different pack sizes are available, providing sufficient material to prepare 100, 500 or 10,000 templates.

Reader Enquiry No. 108

Ambion's NorthernMax is a complete **northern kit** containing all reagents and solutions necessary for detection and quantitation of mRNA. The kit has been designed for use with the company's BrightStar chemiluminescent labelling and detection system, which is said to generate high signal-to-noise ratios and low background, with sensitivities stated to be as good or better than radioisotopic analysis. According to the manufacturer, the kit yields improved results in an isotopic format in minutes instead of days.

ADVERTISEMENTS

£45 per reaction

- PCR products
- M13 inserts
- plasmid inserts
- single strand DNA
- double strand DNA
- no result? no charge!

By research labs for research labs

Molecular Biology Services
 King's College School of Medicine and Dentistry,
 Rayne Institute, 124 Colindale Avenue, London NW9 5EQ
 Tel: 0171-345 3125 Fax: 0171-732 0877

READER ENQUIRY NO 11

NEWCASTLE UNIVERSITY
MOLECULAR BIOLOGY UNIT

PROTEIN SEQUENCING
Membrane-bound, solid or solution samples
£12 per cycle
 5 cycles minimum - no set up charge

For further information contact Dr. Joe Gray
 Phone: +44 191 222 8612
 Fax: +44 191 222 8129
 Email: JOE.GRAY@newcastle.ac.uk

READER ENQUIRY NO 18

Reagents provided with the kit include agarose, denaturing gel buffer, gel loading buffer, gel running buffer, one-hour transfer buffer, high-sensitivity pre-hybridization/hybridization buffer, low- and high-stringency wash buffers, RNase Zap decontamination solution, positive-control target and probe template, and a comprehensive manual.

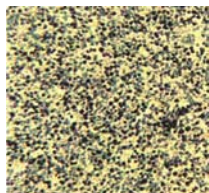
Reader Enquiry No. 109

Reagent register

Promega's 1-kb DNA ladder is designed for **size determinations of dsDNA** from 250 to 10,000 base pairs. The ladder consists of 13 blunt-end fragments ranging in length from 250 to 10,000 bp in evenly-spaced increments. Easy identification of the 1,000- and 3,000-bp fragments is said to be ensured with the increased intensity of these bands. A separate tube of blue/orange 6 × loading dye is supplied with the product. The ladder may be optionally labelled with radioisotopes using T4 polynucleotide kinase and is quality tested for band location, sharpness, relative intensity and absence of background.

Reader Enquiry No. 110

Life Technologies has introduced its DMRIE-C reagent, the newest reagent in the Gibco BRL **cationic lipid reagent** family. This Reagent is designed to improve the transfection of suspension cell lines, such as Jurkat, K562, KG-1 and MOLT-4, as well as for transfection of RNA into cell lines, such as BHK-21, CDS-7, and CHO-K1.



Gibco BRL improves transfection with cationic lipid reagent.

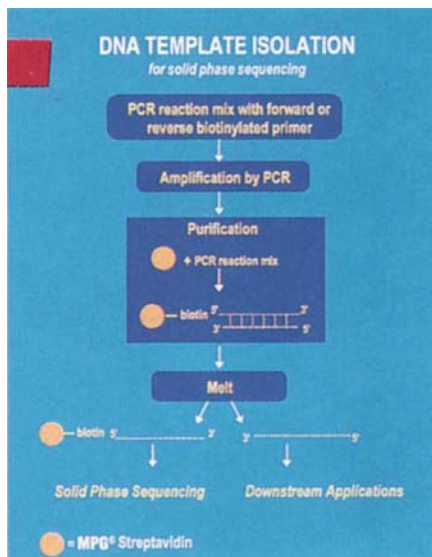
Reader Enquiry No. 111

Quantum Biotechnologies focuses on the development of **research reagents for molecular biology**, genetic engineering and recombinant technologies. The company has introduced the Quant-Essential T7 polymerase *in vitro* mutagenesis kit to allow site-directed mutagenesis to be performed directly on dsDNA purified on a CsCl gradient or by an alkaline lysis, miniprep procedure. These T7 polymerase products are currently available in a user-friendly kit format or as individual components. The company also plans to offer the Klenow kit in the near future.

Reader Enquiry No. 112

Purification

Purification and sequencing of PCR fragments up to 3.5 kb can be accomplished quickly and accurately using CPG's **streptavidin magnetic-porous-glass (MPG)**



PCR-fragment purification and separation.

particles. After PCR with biotinylated primers, the amplified 5'-biotinylated PCR fragments are captured on MPG-streptavidin. Magnetic separation allows the efficient removal of unincorporated dNTPs, enzymes and buffers. Complimentary strands are melted away leaving the single-stranded, immobilized template ready for sequencing. This method is said to eliminate the need for lengthy purification steps requiring spin columns, gels and precipitations. These templates are compatible with either fluorescent or isotopic labelling techniques and have been tested using the AutoRead sequencing kit, Sequitherm long-read cycle-sequencing kit, as well as the Sequenase version 2.0 DNA sequencing kit. The MPG streptavidin-immobilized templates generate clean read lengths reaching 700–800 bases from either manual or automated sequencing. Reproducible results are obtained within a wide usage range (100 µg to 600 µg MPG-streptavidin) by either manual or automated sequencing.

Reader Enquiry No. 113

Integrated Separation Systems has announced the release of the AutoGen P850, an addition to its line of fully **automated systems for isolation and purification of DNA**. The P850 is designed for the customer who only needs to run 50 samples a day. It can be used with DNA from a variety of sources, including plasmids, cosmids, BACs and YACs. The instrument should provide increased productivity by handling higher throughput without exposure to toxic agents or time-consuming manual procedures. The fully automated, microprocessor-controlled system features reagent delivery, robotic tube transfer and a centrifuge module for consistent sample processing. Up to forty-eight samples can be processed in about 5 hours. The AutoGen P850 should prove useful for lab-

oratories performing restriction nuclease digestions, automated fluorescent sequencing, PCR and Southern blotting. It comes with a start-up kit of reagents and consumables for 480 samples. Also available is the AutoGen NA1000, a fully automated system for isolating genomic DNA from animal cells. This floor-model unit is a microprocessor-controlled system, featuring reagent delivery, robotic tube-transfer mechanisms, a reciprocal shaker, a heater module for temperature control and a centrifuge module. Up to forty samples can be processed simultaneously with processing times from 5.5–7.6 hours, depending on the materials being purified. The system should prove useful for genomic isolations from whole blood, lymphocytes, cultured cells and animal tissues, and comes with enough reagents and consumables for 400 samples.

Reader Enquiry No. 114

Invitrogen's Snap **miniprep kit** uses a plasmid binding resin that is said to allow purification of plasmid DNA in less than 30 min. Each lot of miniprep kits is tested in an automated sequencing reaction to ensure a minimum of 450 bases of readable sequence with stated read lengths as high as 650 base pairs. Plasmid DNA is eluted from the resin at the optimal concentration recommended in automated sequencing protocols. The plasmid DNA is also suitable for other molecular biology applications, including mammalian transfection.

Reader Enquiry No. 115

ADVERTISEMENTS

TaqEXPRESS

Nuclease deleted Taq DNA Polymerase (KlenTaq)
(e.g. PNAS 91 pp2216-2220)

- Greatly superior performance in primer extension reactions
- Stable at 99°C
- Patented and licensed (US 5,436,149)
- Inexpensive: £0.09/unit (sold in 500 and 5000 units)

To order: info@genpak.co.uk
Fax: +44 1273 704499 Tel: +44 1273 704470

GENPAK Limited

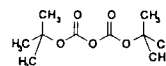
Science Park Square, Falmouth, Cornwall PL11 9SB, UK

READER ENQUIRY NO 53

Now available in ton lots from EU production:

Di-tert-butylidicarbonate

Protecting agent for peptide synthesis



Distributed by:



Produced by:
CHEMTEC LEUNA
Geleleite für Chemie
und Technologie mbH

SIBER HEGNER RAW MATERIALS LTD.
Wiesenstrasse 8, P.O. Box 888 Phone: +41-1-386 73 36
CH-8034 Zurich, Switzerland Fax: +41-1-383 16 63

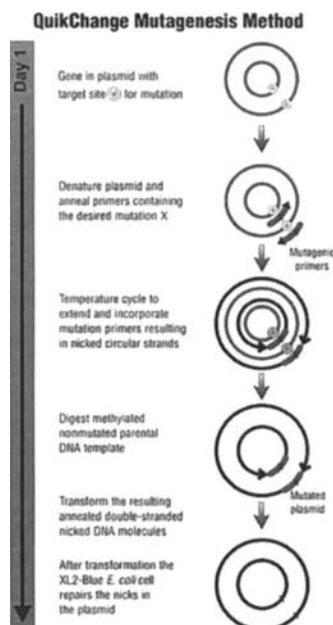
READER ENQUIRY NO 2

Mutation methods

Genosys has just introduced its new p53 GeneScout kit for rapid detection of p53 gene mutations in human cancers. The kit is designed for the analysis of mutations in the human p53 gene coding sequence, allowing PCR amplification of exons five through nine of the p53 gene. After amplification, the PCR product is captured with a capture device that fits onto a laboratory pipette. Following a simple strand-separation step, the template can be sequenced with the primers directed at exons 5, 6, 7, 8 or 9, which are included. Sense and antisense strands of the gene may be isolated separately for sequence analysis. Moreover, the primers have been designed to function in both DNA sequencing and in PCR amplification for single-stranded conformation polymorphism (SSCP) analysis. The kit includes a detailed protocol, necessary primers and additional reagents. Mapping the p53 gene mutations has important implications for accurate diagnosis and compilation of cancer prognostic information. The manufacturer states that more than half of all human malignancies diagnosed each year have a mutated or missing p53 allele. The p53 tumour suppressor gene encodes a DNA-binding protein, and the most prevalent mutations are found within the DNA binding domain encoded by the conserved exons four to nine. These alterations are very likely to affect normal function of the p53 protein, and may lead to uncontrolled cell growth.

Reader Enquiry No. 116

Stratagene's QuikChange site-directed mutagenesis kit is designed to provide an easy and efficient alternative to traditional site-directed mutagenesis techniques. Site directed mutagenesis replicates without PCR, generating site-specific mutations in virtually any double-stranded plasmid and eliminating the need for ssDNA rescue, subcloning into specialized vectors, unique restriction sites or multiple transformations. This mutagenesis method uses a simple one-day, four-step protocol. Only a small starting amount of either purified CsCl or miniprep plasmid DNA is needed and replication will result in increased amounts of DNA. Unlike PCR-based mutagenesis, this method relies on linear extension of template DNA by using high-fidelity *Pfu* DNA polymerase. Mutant oligonucleotide primers, each complementary to opposite strands of the vector, are extended during temperature cycling by the action of *Pfu* polymerase. Due to the unique primer design, only copies of the parental plasmid incorporating the mutation of interest are replicated. This is said to reduce second-site mutations common with



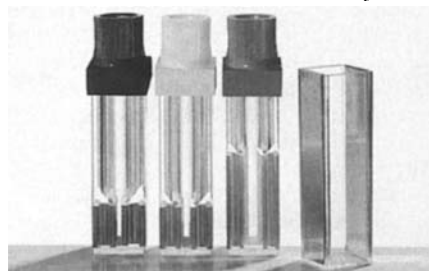
Stratagene site-directed mutagenesis kit.

classic PCR amplification. *Dpn* I treatment after cycling ensures destruction of only the *dam*-methylated parental strands. The resulting nicked, circular mutagenic strands are transformed into *Escherichia coli*, where bacterial ligase repairs the nick and allows plasmid replication. After plating, 80–100 per cent of the colonies contain the desired mutation. The small amount of starting DNA required to perform mutagenesis, the low error rate of *Pfu* DNA polymerase and the low cycle number all contribute to the kit's high mutagenesis efficiency and decreased potential for random mutations.

Reader Enquiry No. 117

Laboratory assistance

EquiBio has launched a new range of electroporation cuvettes. The 'can' design should allow easier aseptic handling and minimize aerosol contamination. The cuvette has been reengineered to reduce dead volumes while maintaining precision and the use of high-quality biotested materials. In practical terms, the user should see less cross-contamination and improved yields. In addition, EquiBio has launched a 10-mm cuvette that is speci-



EquiBio's new electroporation cuvettes.

cally designed for the electroporation of embryos, plant tissue, plant buds and other large cells, such as fish eggs.

Reader Enquiry No. 118

Takasago Electric has reengineered its product line of miniature sterile valves and regulators, which, according to the manufacturer, represent significant size reductions. The JTV-2 valve, for example, is only 34 mm long and just 14.5 mm in diameter. It is DC-solenoid operated for reliability and uses a diaphragm construction so that aggressive chemicals (for example, TFA) are exposed to only the polyarylated resin body and the FPM fluororubber diaphragm. The sampling volume is minimized by 0.9- μ l inlet and 3.6- μ l outlet internal volumes; the response time is a stated 10 ms. The company has also miniaturized their 0.8 to 1.4 kg cm^2 gas regulators through an 80 per cent reduction in volume, down to just a 25-mm cube and no loss of performance. The NRP-500P pump, for DNA use, has a shot volume that can be manually adjusted from 10 μ l to 500 μ l. The structure of the DC-solenoid and the diaphragm are said to ensure reliability and long life, and gives a shot-volume accuracy of ± 1 per cent from 50 μ l to 500 μ l and 2 per cent from 10 μ l to 50 μ l.

Reader Enquiry No. 119

ThermoQuick PCR plates from Greiner Labortechnik are designed to offer enhanced performance, improved sample capacity and flexibility of use for all applications involving thermal cycling. Manufactured from thin-film polycarbonate, these plates should offer enhanced heat transfer from the thermal cycler block, as well as heat tolerances up to 135 $^{\circ}\text{C}$. The plates should prove to have low protein binding and do not release interfering or inhibitory compounds. The 96-well plates are produced in five different versions to fit the most commonly available thermal cyclers. The 96-well format provides both increased capacity and convenience when multiple sample processing is required. However, for those with smaller sample numbers, a 25-well (5 $\times$ 5) plate is also available. Polycarbonate lids, or any of the commonly available sealing films can be used where additional protection is required.

Reader Enquiry No. 120

Correction

In the review entitled "Doxorubicin in sterically stabilized liposomes" (Nature 380, page 562, middle column) 0.72 g g^{-1} should have read 0.72 mg g^{-1} .

Erratum

The review of "Simplified hot start PCR", Nature 381 445–446, had the reader enquiry number omitted. The reader enquiry number for this article is 124.

These notes are compiled by Brendan Horton from information provided by the manufacturers. For more details, fill in the reader service card bound inside the journal.